Evolution of Human and Non-human Primate CC Chemokine Receptor 5 Gene and mRNA

POTENTIAL ROLES FOR HAPLOTYPE AND mRNA DIVERSITY, DIFFERENTIAL HAPLOTYPE-SPECIFIC TRANSCRIPTIONAL ACTIVITY, AND ALTERED TRANSCRIPTION FACTOR BINDING TO POLYMORPHIC NUCLEOTIDES IN THE PATHOGENESIS OF HIV-1 AND SIMIAN IMMUNODEFICIENCY VIRUS*§


From the §Departments of Medicine, University of Texas Health Science Center at San Antonio and South Texas Veterans Health Care System, Audie L. Murphy Division, San Antonio, Texas 78229, the ¶Department of Pediatrics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, the **Department of Virology and Immunology, Southwest Foundation for Biomedical Research, San Antonio, Texas 78228, and the §§Infectious Diseases Service, Wilford Hall Medical Center, Lackland Air Force Base, San Antonio, Texas 78236

Polymorphisms in CC chemokine receptor 5 (CCR5), the major coreceptor of human immunodeficiency virus 1 (HIV-1) and simian immunodeficiency virus (SIV), have a major influence on HIV-1 transmission and disease progression. The effects of these polymorphisms may, in part, account for the differential pathogenesis of HIV-1 (immunosuppression) and SIV (natural resistance) in humans and non-human primates, respectively. Thus, understanding the genetic basis underlying species-specific responses to HIV-1 and SIV could reveal new anti-HIV-1 therapeutic strategies for humans. To this end, we compared CCR5 structure/evolution and regulation among humans, apes, Old World Monkeys, and New World Monkeys. The evolution of the CCR5 cis-regulatory region versus the open reading frame as well as among different domains of the open reading frame differed from one another. CCR5 cis-regulatory region sequence variation in humans was substantially higher than anticipated. Based on this variation, CCR5 haplotypes could be organized into seven evolutionarily distinct human haplogroups (HH) that we designated HHA, -B, -C, -D, -E, -F, and -G. HHA haplotypes were defined as ancestral to all other haplotypes by comparison to the CCR5 haplotypes of non-human primates. Different human and non-human primate CCR5 haplotypes were associated with differential transcriptional regulation, and various polymorphisms resulted in modified DNA-nuclear protein interactions, including altered binding of members of the NF-κB family of transcription factors. We identified novel CCR5 untranslated mRNA sequences that were conserved in human and non-human primates. In some primates, mutations at exon-intron boundaries caused loss of expression of selected CCR5 mRNA isoforms or production of novel mRNA isoforms. Collectively, these findings suggest that the response to HIV-1 and SIV infection in primates may have been driven, in part, by evolution of the elements controlling CCR5 transcription and translation.

Simian immunodeficiency viruses (SIVs)1 comprise a large and genetically diverse group of lentiviruses that originated in sub-Saharan Africa (1–3). SIVs isolated from chimpanzees and mangabeys are very similar to human immunodeficiency virus (HIV)-1 and HIV-2, respectively (3–7). This similarity suggests that HIVs arose via cross-species transmission from non-human primate viral reservoirs. Yet, despite their common ancestry and close similarity, HIVs and SIVs differ significantly with regard to clinical disease and pathogenesis. Human infection with HIVs results in a progressive immunodeficiency syndrome, whereas African apes and monkeys (e.g., African Green monkeys (AGM)) infected with SIV exhibit no evidence of disease (7–9). These differences in pathogenicity may be due, in part, to primate species-specific variation in the genes controlling the host response or expression of host HIV/SIV entry factors (10). Thus, understanding the evolution and factors

* This work was supported in part by a Veterans Administration Merit award, a Veterans Administration Hispanic Research Initiative award (to S. K. A.), by National Institutes of Health Grants RO1-AI42279 and R21-AI46326 (to S. K. A.), RO1-AI14396 (to J. S. A.), and RO1-RR-08122-07 (to K. K. M.), by National Science Foundation Grants SBR-9514733, SBR-9700729, and SBR-9818215, and by National Institutes of Health Grant GM-52290 (to M. B.). This work was also supported by a Robert J. and Helen C. Kleberg Foundation Grant (to S. S. A. and S. K. A.) The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this article (available at http://www.jbc.org) contains Figs. S1–S8.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF246899-AF246925 and AF252551-AF252598.

†† Both authors contributed equally to this work.

¶ Supported by the Robert J. and Helen C. Kleberg Foundation.

¶¶ To whom correspondence should be addressed. E-mail: shujas@uthscsa.edu.

§§ Supported by a Veterans Affairs Career Development Award.

1 The abbreviations used are: SIV, simian immunodeficiency virus; HIV-1, human immunodeficiency virus 1; AGM, African Green monkey; CCR, CC chemokine receptor; UTR, untranslated region; SNP, single nucleotide polymorphism; OWM, Old World monkey; NWM, New World monkey; HH, human haplogroup; PTM, Pig-tailed macaque; NF, nuclear factor; ORF, open reading frame; bp, base pair; PCR, polymerase chain reaction; RT-PCR, reverse transcription-PCR; RACE, rapid amplification of cDNA ends; PBMC, peripheral blood mononuclear cells; EMSA, electrophoretic mobility shift assay; IL, interleukin; HEK, human embryonic kidney.
that control the expression of these genes in primates will be an important step towards identifying the molecular mechanisms underlying the response of primates to infections with SIVs and HIVs. In turn, this may illuminate potential strategies that could be used to mitigate or prevent infection with HIV-1.

An important host genetic determinant of HIV-1 pathogenesis is polymorphisms in the cis-regulatory region of CC chemokine receptor 5 (CCR5), a major coreceptor for the entry of HIV and SIV (reviewed in Refs. 10 and 11). These polymorphisms could potentially influence cell-surface density of CCR5 and thus have an impact on HIV/SIV pathogenesis. For example, homozygosity for a 32-bp deletion in CCR5 ORF leads to loss of surface expression and profound resistance against HIV-1 infection (12). Similarly, a 24-bp deletion in the CCR5 ORF that was discovered in selected non-human primates might influence SIV pathogenesis (13). More recently, we and others (14–17) have shown that polymorphisms in the cis-regulatory region of human CCR5 are also associated with altered rates of disease progression and transmission. Thus, due to this close interaction with the lentiviral life cycle, CCR5 is an excellent candidate for exploring whether differences in its gene/RNA structure and regulation account, in part, for the differential pathogenesis of HIV and SIV. Additional interest in analyzing host-specific differences in CCR5 regulation at the gene and RNA level is spurred by the observation that despite having nearly identical CCR5 coding regions, Asian macaques (e.g. Pig-tailed macaques (PTM)) but not AGMs infected experimentally with SIV become symptomatic (reviewed in Refs. 1, 18, and 19). Given that the differential pathogenicity of AGMs and PTMs to SIV is very reminiscent of the differential pathogenicity of HIV and SIV, studies aimed at dissecting the genetic basis for this differential pathogenicity might provide additional insights into the varied susceptibility to HIV-1 infection among humans.

The gene and RNA structure of human CCR5 is complex. We have demonstrated that alternative splicing in the 5′-untranslated regions (UTR) of CCR5 generates several distinct mRNA isoforms that are under the control of at least two distinct promoters (20). Furthermore, the 5′-UTR of CCR5 is encompassed within these cis-regulatory regions that contain several single nucleotide polymorphisms (SNPs) associated with altered rates of HIV-1 disease progression (14–17). Thus, polymorphisms in the noncoding region of CCR5 could influence not only cis-trans interactions that impact on gene expression but also CCR5 mRNA stability and/or the efficiency of translation.

Given the multiple levels at which CCR5 expression could be regulated, we performed a comprehensive analysis of the ORF, mRNA structure, and transcriptional regulatory units of CCR5 relative to four important events in human evolution (21): the divergence of humans from chimpanzees 6 million years ago, from the Orang-Utan lineage 15 million years ago, from the cercopithecoids (Old World monkeys (OWM)) —35 million years ago, and from New World monkeys (NWM) 50 million years ago. Results from these analyses enabled us to build the evolutionary framework needed to define the relationships among human CCR5 haplotypes that influence HIV-1 pathogenesis. This evolutionary framework facilitated greatly our ability to determine the influence of CCR5 haplotypes on HIV-1 transmission and disease progression (16).2 Additionally, we were able to test directly the hypothesis that polymorphisms in the human and non-human primate cis-regulatory region of CCR5 confer differences in transcriptional efficiencies and/or interact with different trans-acting factors or to the same transcription factor but with varying avidity.

---

2 A. Mangano et al., submitted for publication.

---

**EXPERIMENTAL PROCEDURES**

**Primate CCR5 ORFs**—The CCR5 ORF was PCR-amplified with primers that flanked the human CCR5 ORF (gen gec gec tat gca cgt gga cca ag (forward) and tct agc cta ctt gac tgc tga (reverse)), and cloned, and sequenced on both strands from the following species (common names are in parentheses): Chlorocebus (Cercopithecus) aethiops sabaeus (sabaeus monkey); Hylabates agilis undo (Agile gibbon); Colobus guereza (marmoset); Alestites geoffroyi (black-handed spider monkey); and Lagotrichia lagothricha (woolly monkey). In addition, the following primate sequences in GenBank24 (common names and GenBank accession numbers are in parentheses) were used to construct the CCR5 ORF network (Fig. 1a): Homo sapiens (human, X91492); Pan troglodytes (chimpanzee, AF035214); Gorilla gorilla (gorilla, X54167); Macaca fascicularis (cynomolgus or rhesus monkey, AF075446); Papio hamadryas leucogenys (white-cheeked gibbon, AF075451); Cercocebus torquatus atys (sooty mangabey, AF051905); Cercocebus aterrimus (black mangabey, AF081578); Cercocebus galeritus (Tana River mangabey, AF035215); Macaca fascicularis (crab-eating macaque, AF005660); Macaca mulatta (Rhesus macaque; AF005662); Macaca assamensis (Assamese macaque, AF075449); Macaca arctoides (pig-tailed macaque; AF105282); Papio hamadryas hamadryas (baboon; AF005658); Papio hamadryas anubis (Olive baboon, AF023452); Colobus guereza (black and white colobus, AF114639); Cercopithecus neglectus (De Brazza’s gibbon, AF035218); Cercopithecus nictitans (greater spot-nosed gibbon, AF035219); Cercopithecus 1hoesti (1Hoest’s monkey; AF081579); Cercopithecus cephus (mustached gibbon; AF075440); Saimiri sciureus (red-tailed squirrel gibbon, AF035220); Pygathrix nemaeus (red-shanked douc langur, AF075448); Pygathrix bieli (black snub-nosed monkey, AF075445); Rhinopithecus acentulus (Tonkin snub-nosed monkey, AF075447); Rhinopithecus roxellana (Golden snub-nosed monkey, AF075444); Trachypithecus fridmani ( Francois langur, AF075442); Erythrocebus patas (Patas monkey, AF035220); Chlorocebus aethiops (grivet; AF015944); Chlorocebus aethiops sabaeus (AF035221); Chlorocebus aethiops pygerythrus (vervet; AF035222); Chlorocebus aethiops tantalus (tantalus monkey; AF081577); and Presbytis phayrei (Phayre’s leaf monkey, AF075443). All PCR amplifications were performed with the following conditions: initial denaturation of the DNA template for 90 s, followed by 30 cycles of a profile consisting of 94 °C for 10 s, 55 °C for 30 s, and 72 °C for 1 min. To reduce the potential for errors, a proofreading polymerase, Pfu (Stratagene, La Jolla, CA), was included with the Taq DNA polymerase (Life Technologies, Inc.) at a 1:5 ratio during the PCR amplification.

**CCR5 Numbering System, RNA and Promoter Nomenclature**—The CCR5 numbering system that we used previously (15, 16, 20) was based on the sequences deposited in GenBank24 (accession numbers AF031236 and AF031237) and considered the first nucleotide of the 5′-untranslated region (5′-UTR) to be start site as 1. However, since we had identified numerous 5′-UTR sequences (Fig. 2), and to maintain uniformity in the numbering system used by different investigators, it was proposed at the CCR5-AIDS symposium1 to designate the first nucleotide of the CCR5 translational start site as +1 and the nucleotide immediately upstream as −1 (22). It was also proposed at this meeting to change the numbering system for CCR5 noncoding exons that we had identified previously (20). Exons 2 and 3 are two noncoding exons that are not interrupted by an intron and are now designated as exons 2A and 2B; exon 4 is now designated as exon 3. Finally, to standardize the nomenclature of the CCR5 promoters that we have identified previously (20), in this paper we have designated the downstream CCR5 promoter as promoter 1 and the upstream promoter as promoter 2. A similar nomenclature has been used by other investigators that have multiple promoters (23, 24).

**Primate CCR5 cis-Regulatory Region**—The region corresponding to human CCR5−2761 to −1835 was PCR-amplified, cloned, and sequenced on both strands from the following primates (the number of different members of the given non-human species that were sequenced are shown in parentheses): P. troglodytes (n = 4); G. gorilla; P. pygerythrus (n = 3); M. mulatta (n = 2); M. fascicularis; L. nemesia; C. torquatus (red-capped mangabey); C. galeritus chrysogaster (gold-bellied mangabey); C. guereza; C. guereza kikuyuensis (Kikuyu colobus); Cercopithecus petaurista (lesser spotted gibbon); C. neglectus; C. diana (Diana gibbon); C. 1hoesti; C. (Mopithoepe) talapoin (Talapoin); C. (Erythrocebus) patas; C. aethiops (grivet; n = 3); C. sabaeus (n = 8); C. pygerythrus (n = 3); Presbytis tigrina; Sauromatum genevensis (common two-toed tamarin); C. jacchus; Aotus trivirgatus (owl monkey); A. Geoffroyi; and L.2 A. Mangano et al., submitted for publication.

---

The CCR5-AIDS Symposium was held April 20, 1999, at the NCI-Frederick Cancer Research and Development Center, Frederick, MD.
logstricha. A single allele per non-human primate was sequenced. For Homo sapiens, 60 alleles were sequenced, and they were derived from individuals with different genotypes, including those who were homozygous or heterozygous for −2733A or −2733G, −2135C or −2135T, −1835T or −1835C (15). The CCR5 promoter region from non-human primates was PCR-amplified using the following primers: cat aaa gaa cct gaa ctt gac c (forward) and tag aat ttc taa taa att att cta tta tac tga tga acc aac aac aag gtt cta (reverse). The PCR protocol used to amplify the human and non-human primate CCR5 cis-regions and the ORF was identical. Notably, several alleles from a given species had identical sequences, and repeat amplification/sequencing of the identical allele gave identical results.

**Genotype Analysis of Non-human Primates**—Genotyping methods for the CCR5 polymorphisms were described previously (15, 16). 5′-RACE and Reverse Transcription-PCR (RT-PCR)—Total RNA from human and non-human primate peripheral blood mononuclear cells (PBMC) and human leukocyte subsets was extracted using Trizol reagent (Life Technologies, Inc.). 5′-Rapid amplification of cDNA ends (RACE) was performed on a human leukocyte cDNA library (CLONTECH, Palo Alto, CA) using an exon 2B-specific primer (ggg aac gga tgt ctc acct gta tct cct gtt gct ctt gta ctt). All RT-PCR reactions were run with a negative control that included no cDNA template.

**Promoter Analysis**—To study the differences between the CCR5 promoter activity of sabaeus AGM and that of humans, we constructed a series of firefly luciferase-sabaeus (S1 to S5) and human (H1 to H5) CCR5 promoter constructs in the promoterless pGL3Basic vector (Promega, Madison, WI). A single sabaeus haplotype and a haplotype representative of CCR5 human haplogroup (HHI-A (16) were used to construct the reporter plasmids. The constructs were transfected into human embryonic kidney (HEK), human erythroleukemia (K562), and AGM kidney (COS) cell lines and tested for luciferase reporter activity, as described previously (20). To study differences in promoter activity exhibited by the cis-regulatory regions of human CCR5 haplotypes, the genomic region spanning −2761 to −1814 was PCR-amplified from haplotypes corresponding to HHA, HHC, HHE, HHF, or HHG (16) and cloned into the pGL3Basic vector. Transfection into K562 and Jurkat cell lines and dual luciferase assays were as described previously (20). For all promoter activity analyses, at least two different plasmid preparations were used, and the DNA in each plasmid preparation was quantified spectrophotometrically twice. The Wilcoxon signed-ranks test was used to compare the mean luciferase activity between homologous sabaeus and human promoter constructs. Statistical analysis to determine the differences in the mean luciferase activity among human CCR5 promoter alleles was by one-way analysis of variance followed by the Scheffe’s post hoc test.

**Electrophoretic Mobility Shift Assay (EMSA)**—All cell lines were obtained from American Type Culture Collection and were maintained as described previously (20). Nuclear extracts were prepared from K562, THP-1 (human monocyte), Jurkat (human T-cells), and COS cell lines, according to standard protocols (25). Nuclear proteins from PHA-stimulated human PBMCs were extracted according to the procedure of Schreiber et al. (26). Prior to nuclear protein extraction, human PBMCs were stimulated for 3 h with phorbol myristic acid (25 ng/ml) and ionomycin (1 μg/ml). PBMCs from SIV-uninfected sabaeus AGM and vervet AGM were stimulated with 10 μg/ml PHA for 2 days and were maintained in culture for 11 days in the presence of 10% human IL-2 (Advanced Biotechnologies Inc., Columbia, MD) before extracting the nuclear proteins. Nuclear extracts from similarly PHA/IL-2-treated human PBMCs served as controls for the EMSA experiments that used sabaeus/vervet PHA blasts as nuclear extracts. The nucleotide sequences of the oligonucleotides used in EMSAs are shown in Table I. For competition experiments, unlabeled competitor oligonucleotides were incubated with the nuclear extracts for 10 min on ice prior to adding the labeled probe. The specificity of the binding reactions was confirmed by using nonspecific double-stranded oligonucleotide competitors. For super-shift experiments, the indicated antibodies (Santa Cruz Biotechnology, Santa Cruz, CA) were incubated with the nuclear extracts and the radiolabeled probes for 45 min at room temperature prior to electrophoretic resolution of the complexes. Densitometry analysis of the EMSA gels was performed with the NIH Image software package (version 1.61).

**Phylogenetic Analysis**—Sequences were aligned using the SEQUENCHER software package. Descriptive statistics were obtained using the ARLEQUIN software package. Mean nucleotide diversity within populations was estimated using the equation, π = (n/n−1)Σx−x̄2/n, where n is the number of DNA sequences examined, x, and x̄ are the population frequencies of the i'th and j'th type of DNA sequences, and πij is the proportion of nucleotides that differ between the i'th and j'th types of DNA sequence. Genetic distances between sequences were estimated using DNADIST of the PHYLIP software package using Kimura’s two-parameter model. The transition to transversion ratio was varied from 2:1 to 10,000:1 but had no substantial impact on the results. Distances between populations were estimated from distances between individuals using NEIDIST (27). Relationships between lineages and/or populations were depicted as neighbor-joining networks (28), using NEIGHBOR. Inferred branch lengths with negative values were converted to branches of length zero. Parsimony networks were constructed using DNAPARS. The robustness of branches was assessed by using 100 bootstrap data sets obtained using SEQBOOT. Neighbor-joining and parsimony trees were produced using mouse sequence as an outgroup, and CONSENSE was used to find the consensus tree. Networks were

---

Table I

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Nucleotide sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR5 – 2733A=G</td>
<td>atc tgt aag gaa (A/G) a tcc tgc cac</td>
</tr>
<tr>
<td>CCR5 – 2554G&gt;T</td>
<td>ttt cgg ttt tac gaa</td>
</tr>
<tr>
<td>CCR5 – 2459G&gt;A</td>
<td>gtt gaa aag ggg (G/A) cag agg aat gtg</td>
</tr>
<tr>
<td>CCR5 – 2135T=C and −2132C&gt;T</td>
<td>age cgg taa aa (C/T) tt (G/T) aga cca gac atat</td>
</tr>
<tr>
<td>CCR5 – 2096A&gt;G</td>
<td>aag cca cca tag aag a (G/C) att tgg cca aca</td>
</tr>
<tr>
<td>CCR5 – 1835C&gt;T</td>
<td>ttt cca aag cag gtt ttt gtt tac gaa</td>
</tr>
<tr>
<td>CCR5 OWN – 2554G</td>
<td>ttc cag gaa gtt gtt</td>
</tr>
<tr>
<td>CCR5 NWM-1 – 2554G</td>
<td>gtt</td>
</tr>
<tr>
<td>CCR5 NWM-2 – 2554G</td>
<td>ttc cag gaa gtt gtt</td>
</tr>
<tr>
<td>NF-aD consensus</td>
<td>gta cag gaa ccc gcc ggg</td>
</tr>
<tr>
<td>SP</td>
<td>gaa gga aat gaa aag gaa (A/G) cag agg aat gtg</td>
</tr>
<tr>
<td>IRF-1</td>
<td>age cgg taa aa (C/T) tt (G/T) aga cca gac atat</td>
</tr>
<tr>
<td>PEA3/ETS-1</td>
<td>ttt cca aag cag gtt ttt gtt tac gaa</td>
</tr>
<tr>
<td>c-MYB</td>
<td>gtt gaa aag ggg (G/A) cag agg aat gtg</td>
</tr>
<tr>
<td>cBPα</td>
<td>age cgg taa aa (C/T) tt (G/T) aga cca gac atat</td>
</tr>
<tr>
<td>AP-3</td>
<td>ttt cca aag cag gtt ttt gtt tac gaa</td>
</tr>
</tbody>
</table>
visualized using TREETOOL. Estimates of the rates of nonsynonymous (dN) and synonymous (dS) substitutions for all pairwise comparisons were calculated using the method of Nei and Gojobori (29) as implemented in the PAML package (30).

RESULTS

Molecular Evolution of the CCR5 ORF in Primates—Comparison of the complete CCR5 ORF from 37 different primates revealed that the nucleotide sequence and amino acid identity of CCR5 were highly conserved (alignment is shown in Supplemental Figs. 1 and 2; species list is provided under “Experimental Procedures”). Of the variable sites, 184 were SNPs, including 148 transitions and 55 transversions. To quantify the amount of sequence variation found among primates, we estimated the nucleotide diversity of each group. Nucleotide diversity is equivalent to the number of nucleotide differences per site between a random pair of chromosomes drawn from a population. For all primates, the mean nucleotide diversity of the CCR5 ORF was 0.025 (~1 variant in every 40 bp). Levels of total nucleotide diversity differed substantially among hominoids (humans and apes), OWM, and NWM. Nucleotide diversity in hominoids (0.009) and OWM (0.0013) was substantially lower than that found within the total primate group, whereas nucleotide diversity in NWM was the highest of all primate groups (0.032).

In coding regions, mutation and selection are expected to have different effects on nonsynonymous (dn) and synonymous (ds) nucleotide substitutions. Consequently, comparisons of the rate of dN to dS substitutions (dN/dS) can be utilized to explore molecular sequence evolution (31). The neutral theory predicts that despite varying mutation rates between lineages, dN/dS should remain constant among lineages. Thus, variation of dN/dS among lineages is considered evidence against neutrality, whereas dN/dS ratios >1.0 are evidence for positive selection (32).

Pairwise maximum likelihood estimates of dN/dS among primate CCR5 ORFs were consistently <1.0. However, estimation of dN/dS for each of the domains of CCR5 (e.g. NH2 terminus, extracellular loops, and intracellular tail) revealed an interesting trend. Pairwise estimates of dN/dS among hominoids and NWM, for the sequence encoding the NH2 terminus, were consistently >1.0. Pairwise estimates of dN/dS for the second extracellular loop among hominoids and two of three NWM (spider or woolly monkey) were >1.0 (dN/dS between hominoids and marmosets was always <1.0). These findings suggested that the effects of natural selection might vary among specific domains of CCR5 ORF. Moreover, these results indicated that substitutions in the NH2 terminus and second extracellular loop may underlie a selective response to the pathogens after the NWM and Catarrhine split. This was consistent with the finding that the bulk of polymorphisms in the human and non-human CCR5 ORF have been found primarily in the NH2 terminus, and only a few are found elsewhere (33, 34).

Phylogenetic reconstruction of the genetic affinities among hominoids, OWM, and NWM demonstrated that NWM were substantially more divergent from either hominoids or OWM (Fig. 1a). That is, the genetic distance between NWM and hominoids (0.054) or OWM and NWM (0.055) was much larger than the genetic distance between hominoids and OWM (0.014). These findings were consistent with estimates of genetic divergence among these groups based upon analysis of morphological and neutral genetic markers (35). Thus, it is noteworthy that despite the different roles that CCR5 may have played in mediating responses to pathogens (e.g. SIV and HIV-1) among African OWM, Asian OWM, and hominoids, the pattern of variation observed among primate CCR5 ORFs is similar to that observed for neutral markers. Overall these data suggest that the expression/function of CCR5 among OWM and hominoids is more likely to be controlled by factors that regulate CCR5 transcription, mRNA processing, and/or translation than selection for different ORF variants. For this reason, we explored the nature of variation in the mRNA structure and cis-regulatory region of CCR5 in NWM, OWM, and hominoids.

Evolution of a cis-Regulatory Region of CCR5 in Non-human Primates—Alignment of the nucleotide sequence of a cis-regulatory region of CCR5 from non-human primates revealed high sequence conservation (Fig. 1b and Supplemental Fig. 3). Nevertheless, substantial intra- and interspecies sequence variation was observed (Fig. 1, c–i). Compared with the human sequence, one gap was required to align the sequence of the chimpanzee CCR5 cis-regulatory region, and six gaps were inserted to align the OWM sequences. No gaps were required to align the gorilla and orang-Utan CCR5 promoter sequences; however, several gaps were required to align the 44 primate sequences (Supplemental Fig. 3).

Compared with the CCR5 ORF, a cis-regulatory region of CCR5 demonstrated similar levels of nucleotide sequence diversity. Of the polymorphic sites, 237 were SNPs, including 177 transitions and 68 transversions. For all primates, the mean nucleotide diversity of the cis-regulatory region of CCR5 was 0.022, which is approximately 1 variant in every 45 bp. Mean nucleotide diversity was 0.007, 0.007, and 0.028 in hominoids, OWM, and NWM, respectively. The cis-regulatory regions of CCR5 in chimpanzee and human differed at 41 sites, including 8 fixed and 33 variable sites, suggesting that the CCR5 cis-regulatory regions of chimpanzees and humans were considerably more different from one another than their CCR5 ORFs.

Genetic distances estimated from the cis-regulatory region of CCR5 of hominoids, OWM, and NWM indicated that hominoids were nearly equally divergent from OWM and NWM (0.058) was comparable to the genetic distance between hominoids and NWM (0.067). This was in contrast to the closer affinity of hominoids and OWM, as estimated from analysis of the CCR5 ORF. In other words, the genetic distance between OWM and NWM was similar regardless of whether the CCR5 ORF or cis-regulatory regions were compared, whereas the genetic distances between OWM and hominoids estimated from the cis-regulatory region of CCR5 was 4 times greater than estimates of the genetic distance from the CCR5 ORF. These data suggested that the CCR5 cis-regulatory region of hominoids was substantially more divergent from OWM than is the CCR5 ORF. This underscores the potential role that different evolutionary forces (e.g. natural selection) may have played in shaping the genetic variation of the cis-regulatory region of CCR5 in hominoids versus OWM.

CCR5 mRNA Splicing Patterns in Primates: mRNA Isoform Diversity—We designed experiments to investigate whether the complex mRNA structure of CCR5 (20) is conserved across different primate species. Furthermore, because the cis-regulatory region of CCR5 encompasses the 5′-UTR region of human CCR5 (Fig. 2a and b), we determined whether the genetic variation we observed in this region in non-human primates would influence the nature of the mRNA isoforms produced in different primates.

We have shown previously that two “full-length” CCR5 mRNA transcripts (CCR5A and CCR5B) arise by alternative splicing (20) (Fig. 2c). Several “truncated” transcripts can also originate in either exon 2A (old nomenclature exon 2) or exon 2B (old nomenclature exon 3) of CCR5 (Fig. 2c). By using 5′-RACE on a human leukocyte cDNA library, we extended the known CCR5 mRNA sequence common to both CCR5A and
CCR5 by 142 additional nucleotides (−2903 to −2762; Fig. 2, a and b). Analysis of the genomic DNA sequence immediately upstream of the newly identified mRNA sequence in exon 1 indicated the presence of several potential transcription factor binding sites (Fig. 2b). This new exon 1 sequence was subsequently found in different human leukocyte subsets as well as in mononuclear cells of several non-human primate species (Fig. 2, c and d, and data not shown).

Comparison of the genomic DNA sequence extending from exon 1 to exon 2B among non-human primates (Fig. 1b and Supplemental Fig. 3) and RNA transcripts in mononuclear cells derived from chimpanzees, rhesus macaque, cynomolgus macaque, and sabaeus AGM revealed the following (Fig. 2, c and d). The exon-intron splice donor and acceptor sites were conserved between humans and Orang-Utan, gorilla, langur, and NWM. Mutations in the exon-intron splice acceptor donor sites lead to loss of expression of selective CCR5 mRNA isoforms in different non-human primates. Alternatively, usage of a non-canonical splice donor site in exon 1 of sabaeus resulted in the retention of this complicated RNA organization may have afforded a selective advantage.

Evolution of a cis-Regulatory Region of CCR5 in Humans—Sequence analysis of the cis-regulatory region (−2761 to −1835) of 60 human CCR5 alleles revealed a total of 32 variable sites that define 27 unique human haplotypes (Fig. 3). An
Fig. 2. Conservation of a complex CCR5 mRNA structure in primates. a, genomic organization of human CCR5. Numbered boxes, exons; slashed lines, gaps; Promoter 1 and Promoter 2 indicate the previously designated downstream (Pd) and upstream (Pu) promoter, respectively (20). 5’-RACE extended the previously described exon 1 by an additional 142 bp (−2762 to −2903 which is equivalent to −1 to −142 of GenBank™ accession numbers AF031237 and AF031237). Arrows under the open boxes indicate the location of the sense and antisense primers used in RT-PCR experiments. b, novel CCR5 mRNA sequence (−2903 to −2762) and genomic sequence upstream of human CCR5 exon 1. Exon and intron/cis-regulatory sequences are in uppercase and lowercase, respectively. The novel mRNA sequences identified in this paper are highlighted in red, and the previously identified exon 1 mRNA sequences (20) are in black uppercase. Arrows indicate the start of exon 1 and intron 1. Putative transcription factor-binding sites upstream of exon 1 are overlined with a straight arrow. c, schema of alternative CCR5 mRNA splicing patterns in primates. mRNA isoforms that contain exon 1 are arbitrarily designated as full-length transcripts (CCR5A and CCR5B), and those lacking it are designated as truncated transcripts and are not assigned a name (20). Tick marks indicate exon boundaries, and the angled dashed lines indicate the splicing patterns. gt and ag indicate the splice donor and acceptor sites, respectively. The mutations in the splice acceptor site are denoted in red. These mutations prevent the production of the CCR5 mRNA isoforms that are marked by an X (e.g. CCR5B was not found in cynomolgous, rhesus or sabaeus AGM). In a sabaeus monkey, a CCR5 mRNA species was found whose structure includes a truncated exon 1 (lacks the region corresponding to human CCR5 −2752 to −2705) and exons 2A, 2B, and 3. The non-consensus splice donor site of the truncated sabaeus exon 1 is shown in green. d, expression of novel exon 1 sequence in primates. An ethidium bromide-stained gel of the RT-PCR products is shown. The sources for the RNA used in RT-PCR are indicated on the top, and abbreviations used are as follows: P, human peripheral blood mononuclear cells; L, human lymphocytes; AC, human activated CD4+ cells; and SAB, peripheral blood mononuclear cells from a sabaeus AGM. The location of the final set of primers used in the RT-PCR is shown in a. The length of the products amplified from human cells is indicated on the left. In sabaeus, the isoform homologous to CCR5A is shorter because of a deletion (48 nucleotides) in the 3’ end of the region corresponding to human CCR5 exon 1.
Fig. 2. CCR5 gene map and phylogenetic network of human CCR5 haplotypes and haplogroups. A schematic illustration of the human CCR5 ORF (GenBank accession number AF009962; deposited by Dr. N. L. Michael's group) and five non-human primates (Fig. 1b and Supplemental Fig. 3) and genotypic data from 40 additional non-human primates, including 23 chimpanzees, enabled us to define the CCR5 haplotype ancestral to humans. That is we determined the polarity (the ancestral-descendant relationship) of each nucleotide variant in the cis-regulatory region of human CCR5. In previous studies (15, 20), we found seven common polymorphic sites in the region between CCR5 2761 to 1835 (Fig. 3a). -2733A, -2554G, -2459A, -2135T, -2123C, -2086A, and -1835C represented the ancestral state for these variable sites in human CCR5 (Figs. 1b and 3c). The nucleotide identity at each of these positions was invariant among great apes (except Gorilla which had a CCR5 2132T) and OWM (Fig. 1b). This ancestral CCR5 haplotype was used to root a phylogenetic network depicting the possible evolutionary relationships among unique human CCR5 haplotypes (Fig. 3b).

A phylogenetic network of unique CCR5 haplotypes provided the evolutionary framework for defining seven biologically distinct clusters of haplotypes that we designated as CCR5 human haplogroups (HHA-HHG). HHA represented the ancestral CCR5 haplogroup. The haplogroups, HHC through HHG, were defined by at least one SNP, i.e., SNPs 2086G, 2132T, 1835T, and 2733G distinguish CCR5 haplogroups HHC, HHD, HHF, and HHG, respectively. HBB haplotypes had a 2554T mutation but lacked the 2132T and 2086G SNPs. An HHB haplotype is likely to be ancestral to HHC and HHD (Fig. 3c). SNPs 2459A and 2135C were in complete linkage disequilibrium. Haplotypes with 2459A and 2135C but lacking 2733G and 1835T defined HHE. The polymorphisms CCR5 2733G, 1835T, and 2135C define the haplotypes that are descendants of ancestral haplotypes in HHE (Fig. 3c). The CCR2-64I and CCR5-32 polymorphisms were found only on CCR5 haplotypes in haplogroups F, HHC, HHD, HHF, and HHG, respectively. To assess the robustness of each of the branches that define, in part, human CCR5 haplogroups, a bootstrap analysis was performed. Each branch was observed in 60% or more of the networks generated (Fig. 3b). Collectively, these findings demonstrate that SNPs in CCR5 may have arisen by a nested mutational process and that this locus represents a complex multi-allelic system. A clear understanding of the extent and organization of genetic variation in human CCR5, including the classification system we developed herein, permitted us to investigate the role of CCR5 haplotypes in vertical/horizontal transmission and disease progression in infected HIV-1 adults and children (16).2

Functional Effects of Variation in a cis-Regulatory Region of Human CCR5: Haplotype-specific Differences in Transcriptional Activity—Because of the influential effect that CCR5 haplotypes had on HIV-1 transmission/disease progression (14–17), we next initiated studies to determine the potential mechanisms underlying their effects. To test the hypothesis that polymorphisms in the cis-regions of human CCR5 haplotypes confer haplotype-specific differences in transcriptional activity, we cloned CCR5 2761 to 1814 derived from haplotypes representing several of the CCR5 haplogroups upstream of the luciferase ORF. These constructs represent haplotype-specific CCR5 promoter constructs because they included the cis-region of CCR5 that encompassed the polymorphic sites that demarcated the major haplogroups. We found significant differences among the luciferase activity of the five haplotype-specific promoter constructs tested, with the HHA-specific promoter construct demonstrating the least promoter activity (Fig. 4a).
Functional Effects of Variation in a cis-Regulatory Region of Human CCR5: Altered DNA/Nuclear Factor Binding—Another potential mechanism by which polymorphisms in the cis-region of CCR5 haplotypes could influence CCR5 expression is by binding to distinct nuclear factors or altering the avidity of binding to a transcription factor. This might also provide a reason for the haplotype-specific differences in transcriptional activity. Computer algorithms suggested that the cis-regions surrounding the SNPs that delimit the major human haplogroups could potentially bind to distinct transcription factors (Table II). We used EMSAs and supershift assays to determine whether polymorphisms at −2733, −2554, −2459, −2315, −2132, −2086, or −1835 bound nuclear proteins with differing avidity or whether a particular polymorphism bound a novel transcription factor. Table III summarizes the findings of these studies.

We found that SNPs in human CCR5 cis-regulatory sequences were associated with three different kinds of alterations in DNA/nuclear factor binding. First, gain-of-nuclear factor binding: radiolabeled −2733G but not −2733A oligomers (Table I) bound to a novel nuclear factor (NF1) in three cell human cell lines (Fig. 4b and Table III). When using −2733G, two additional bands were found below NF1 in Jurkat cells, and it was difficult to exclude the possibility that one of these bands corresponded to that observed in EMSAs with −2733A (Fig. 4b). These two bands were also found in EMSAs with the −2733G oligomer in K562 or THP1 cells; however, the intensities of these bands was significantly lower than that found in Jurkat cells (Fig. 4b, compare lane 8 with lanes 6 and 7). Only NF1 is highlighted in Fig. 4b because this gain-of-nuclear factor binding was consistently present in the three cell lines tested.

Second, loss-of-nuclear factor binding: radiolabeled −1835C oligomer bound specifically to two novel nuclear factors (NF2 and NF3), but in contrast the radiolabeled −1835T oligomer bound to only one of these factors (NF2; Fig. 4b and Table III). The radiolabeled −1835C and −1835T oligomers also bound an additional nuclear factor that appears as a band below NF3, and this band is competed weakly by unlabeled homologous or heterologous oligomers but not by a nonspecific oligomer (Fig. 4b). Notably, the altered DNA/NF1-NF3 binding was observed in human cell lines but not in PHA-blast cells, suggesting that this effect may be cell type- and/or cell activation-specific.

Finally, a SNP in human CCR5 was associated with altered avidity of nuclear factor binding. This observation was illustrated by the altered binding pattern of several members of the NF-κB family of transcription factors to the cis-region that encompasses the human CCR5 G−2554T polymorphism (Fig. 5a). Radiolabeled −2554G and −2554T oligomers each bound to four nuclear factor complexes present in nuclear proteins derived from human PHA blasts (NF4-NF7; Fig. 5b). In homologous competition assays, unlabeled −2554G and −2554T oligomers competed for the NF4-NF7 bound to the radiolabeled −2554G or −2554T oligomers, respectively, and the efficiency for competition was as follows: NF6 > NF5 > NF4 > NF3 (Fig. 5b, compare intensity of bands in lanes 2 and 3; and data not shown for −2554G competition). By densitometric analysis, the intensities of the NF5 and NF4 complexes that bound to the −2554G probe were 198 ± 28.13 and 371 ± 42% greater than the intensities of these complexes that bound to the −2554T probe, respectively (n = 5 separate experiments using nuclear extracts from the same donor). Notably, this altered DNA/NF-κB binding was observed in nuclear extracts from human PHA blasts but not from unstimulated Jurkat T-cells, again highlighting that these altered DNA/nuclear protein interactions to human CCR5 SNPs are likely to be cell type- and/or cell activation state-specific.

Because the region encompassing human CCR5 G−2554T was predicted to bind to members of the NF-κB family of transcription factors (Table II and Fig. 5a), we examined the ability of an unlabeled oligomer that contains a canonical NF-κB-binding site to compete for the NF4-NF7 bound to the labeled −2554G or −2554T oligomers. The unlabeled consensus NF-κB-binding oligomer competed for the NF4-NF7 bound to the labeled −2554G or −2554T oligomers, but the efficiency of competition was different than that for unlabeled −2554G/−2554T and was NF5 > NF4 = NF7 > NF6 (Fig. 5b; compare intensity of bands in lanes 4 and 5 to lanes 2 and 3, and data not shown for competition to −2554G).
Two computer algorithms (TESS and MatInspector program) were used to predict the transcription factors that could potentially bind to the SNPs in a cis-regulatory region (~2761 to ~1835) of human CCR5.

<table>
<thead>
<tr>
<th>CCR5 SNP(s)</th>
<th>TESS Predicted transcription factor binding sites</th>
<th>MatInspector</th>
</tr>
</thead>
<tbody>
<tr>
<td>−2733G</td>
<td>SP-1, c-ETS-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>c-ETS</td>
</tr>
<tr>
<td>−2554G</td>
<td>c-ETS-1, c-Rel</td>
<td>c-MYB, PEA-3&lt;sup&gt;c&lt;/sup&gt;, ELK-1&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>−2459G</td>
<td>CAC-binding protein, CF-1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>c-MYB, IRF&lt;sup&gt;aa&lt;/sup&gt;</td>
</tr>
<tr>
<td>−2459A</td>
<td>CAC-binding protein, v-JUN&lt;sup&gt;c&lt;/sup&gt;</td>
<td>MZF, IK2, USF, c/EBPα&lt;sup&gt;aa&lt;/sup&gt;</td>
</tr>
<tr>
<td>−2135C and −2132C</td>
<td>E-12&lt;sup&gt;c&lt;/sup&gt;, v-MYB&lt;sup&gt;a&lt;/sup&gt;, HFH-1</td>
<td></td>
</tr>
<tr>
<td>−2135T and −2132C</td>
<td>HFH-1</td>
<td></td>
</tr>
<tr>
<td>−2135T and −2132T</td>
<td>HFH-1</td>
<td></td>
</tr>
<tr>
<td>−2086A</td>
<td>NFAT, TCF-1&lt;sup&gt;c&lt;/sup&gt;, c-ETS-2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>v-MYB</td>
</tr>
<tr>
<td>−2086G</td>
<td>NFAT, PU-box, c-ETS-1&lt;sup&gt;c&lt;/sup&gt;, PEA-3&lt;sup&gt;c&lt;/sup&gt;</td>
<td>v-MYB</td>
</tr>
<tr>
<td>−1835C</td>
<td>AP3, c/EBPα&lt;sup&gt;a&lt;/sup&gt;, NF4&lt;sup&gt;c&lt;/sup&gt;, MyoD&lt;sup&gt;a&lt;/sup&gt;, E12&lt;sup&gt;c&lt;/sup&gt;</td>
<td>c/EBPβ</td>
</tr>
<tr>
<td>−1835T</td>
<td>AP3, c-ETS-1&lt;sup&gt;c&lt;/sup&gt;, SRF&lt;sup&gt;a&lt;/sup&gt;</td>
<td>c/EBPβ</td>
</tr>
</tbody>
</table>

<sup>a</sup> Denotes that the predicted binding of the transcription factor is unique to the indicated nucleotide. For example, the oligonucleotide that contains −2554T but not −2554G is predicted to bind the IRF transcription factor. The binding of transcription factors to the three possible SNP combinations at CCR5 −2135 and −2132 (Fig. 3c) are shown. Note that only a selected number of the transcription factors predicted to bind to the CCR5 SNPs are shown, and the output from the two computer algorithms used was not always identical.

### Table III

Summary of EMSA findings for cis-regions spanning SNPs at CCR5 −2733, −2554, −2459, −2135, −2132, −2086, and −1835

Oligonucleotides that encompassed different SNPs (Table I) in a cis-regulatory region of CCR5 (~2761 to ~1835) were incubated with nuclear extracts derived from different cell lines or PHA-stimulated human PBMCs (PHA-blasts). Whether a SNP resulted in the differential (diff.) nuclear factor binding and/or altered affinity of nuclear factor binding is indicated. Results of supershift assays and the competition experiments with oligonucleotides containing the consensus binding sequences for the indicated nuclear factors are also shown. The single letters within the parentheses indicate the cells from which the nuclear extracts were prepared and used in EMSAs (J, Jurkat; K, K562; T, THP-1; and P, PHA blast), + indicates that competition or a supershift was present; − indicates that competition or a supershift was absent. NT indicates not tested. "Identical binding" indicates that the nuclear factor binding pattern of the two polymorphic oligonucleotides to nuclear factors in a given cell line or in PHA-blasts was similar. Note, the binding pattern among the cell lines or between the cell lines and the PHA-blasts was not necessarily identical.

<table>
<thead>
<tr>
<th>CCR5 SNP</th>
<th>Cell lines (K, J, T)</th>
<th>PHA-blasts</th>
<th>Competition</th>
<th>Supershift</th>
</tr>
</thead>
<tbody>
<tr>
<td>−2733A&gt;G</td>
<td>Diff. binding to NF1 (Fig. 4b)</td>
<td>No binding to NF1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>SP-1 (J) (−)</td>
<td>NT</td>
</tr>
<tr>
<td>−2554G&gt;T</td>
<td>Identical binding&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Greater avidity of −2554T for binding to NF4–NF7 than −2554G (Fig. 5)</td>
<td>NF-κB (P) (+)</td>
<td>p65 (P) (+)</td>
</tr>
<tr>
<td>−2459G&gt;A</td>
<td>Identical binding&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Identical binding&lt;sup&gt;b&lt;/sup&gt;</td>
<td>NF-κB (P) (+)</td>
<td>p65 (P) (+)</td>
</tr>
<tr>
<td>−2135T&gt;C and −2132C&gt;T</td>
<td>Identical binding&lt;sup&gt;b&lt;/sup&gt;</td>
<td>NT</td>
<td>NF-κB (P) (+)</td>
<td>p65 (P) (+)</td>
</tr>
<tr>
<td>−2086A&gt;G</td>
<td>Diff. binding&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>No binding to NF2 or NF3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NC-κB (K) (−)</td>
<td>E12 (K) (−)</td>
</tr>
<tr>
<td>−1835C&gt;T</td>
<td>Diff. binding&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>Diff. binding&lt;sup&gt;d&lt;/sup&gt;</td>
<td>c/EBPβ (K) (−)</td>
<td>MyoD (K) (−)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Data not shown.
<sup>b</sup> Not tested in THP-1 nuclear extracts.
<sup>c</sup> Not tested in K562 nuclear extracts.
<sup>d</sup> The radiolabeled CCR5 −2086G oligonucleotide bound specifically to a nuclear factor not bound by the −2086A oligonucleotide.

It was not clear from these EMSA competition assays (Fig. 5b) whether each or only a subset of the nuclear complexes that bound to the labeled −2554G and −2554T oligomers, i.e. NF-κB, contained components of the NF-κB/Rel family of nuclear factors. Thus, it was important to compare directly the nuclear protein binding pattern of radiolabeled −2554T or −2554G to that of the radiolabeled consensus NF-κB-binding oligomer (Table I, Fig. 5c). The radiolabeled consensus NF-κB-binding oligomer bound to a complex of factors that appeared to include all or portions of NF4 and NF5 but not NF6 or NF7 (Fig. 5c, compare lanes 1 and 10). In contrast, both radiolabeled −2554T oligomer (Fig. 5c) and −2554G (data similar to Fig. 5b, lane 1) bound NF4–NF7. Notably, the consensus NF-κB-binding oligomer bound to NF-κB family members with greater avidity than the −2554T oligomer (Fig. 5c, compare intensity of bands in lanes 1 and 10). These findings suggested that the cis-regions surrounding human CCR5 G−2554T bound to various members of the NF-κB proteins as well as to two nuclear proteins (NF6-NF7) that are not bound by an oligomer that contains a consensus NF-κB-binding site.

To confirm that NF4 and NF5 but not NF6 or NF7 contain members of the NF-κB family of transcription factors, we conducted supershift assays with antibodies to p65, p50, c-Rel, and p52. By supershift assays, only p50 was detected in the NF4 complex bound to radiolabeled −2554T oligomer, whereas parts of NF5 were supershifted by antibodies to p65, p50, and c-Rel (Fig. 5c, lanes 1–9). Note that the complex supershifted by c-Rel antibody comigrates with NF7. Overall, the pattern of the supershifted bands indicated that NF4 and NF5 bound to the radiolabeled −2554T oligomer contained p65 homodimers, p50/p65 heterodimers, p50 homodimers, and c-Rel/p50 heterodimers (Fig. 5c, lanes 1–9). No supershift was observed with antibodies to p52 or RelB (Fig. 5c, and data not shown). Notably, antibodies to the NF-κB family members did
The CCR5 evolutionary and regulation analysis was performed to understand the binding specificity of the NF-κB family of transcription factors and to explore the functional effects of genetic variations. The CCR5Evolution and Regulation

**Fig. 5.** Differential binding of NF-κB/Rel family members and two as yet unidentified nuclear factors to polymorphisms at human CCR5 −2554. a, the cis-region spanning human CCR5 −2554 and the homologous region in non-human primates share sequence similarities with the consensus binding site for members of the NF-κB/Rel family of transcription factors. CCR5 −2554 is found in HH-A, -E, -F, and -G and non-human primates. −2554T is found in HH-B, -C, and -D. Boxes highlight differences among the different primates. b, EMSA and competition experiments were performed with nuclear extracts from PHA-stimulated human PBMCs with radiolabeled oligomers that contain either −2554G or −2554T (Table I). Both the radiolabeled −2554G (lane 1) and −2554T (lane 2) oligomers bound to four distinct nuclear factor complexes (NF4-NF7); however, the −2554T oligomer bound to these factors with greater avidity than the −2554G oligomer (compare intensity of bands in lanes 1 and 2; see "Results" for densitometric analysis). The unlabeled −2554T oligomer but not an unrelated oligomer competed efficiently and specifically for the binding of NF4-NF7 to the radiolabeled −2554T oligomer (lane 3, and data not shown). The unlabeled −2554G oligomer also competed efficiently and specifically for the binding of radiolabeled −2554G oligomer (data not shown). Lanes 4 and 5 show the competition with excess of the unlabeled consensus NF-κB-binding oligomer. c, supershift experiments to determine the identity of the nuclear proteins bound by the −2554T oligomer in PHA-stimulated human PBMCs. Radiolabeled −2554T or consensus NF-κB oligomers (Table I) were incubated with extracts from human PHA blasts. The complexes were then electrophoretically resolved after addition of the antibodies to the proteins indicated. No interaction between the nuclear protein complexes and an isotype control antibody was observed (data not shown). Note, complexes equivalent to NF6 and NF7 did not bind to the radiolabeled NF-κB oligomer, and the intensity of the NF4 and NF5 bands is greater in lane 10 versus lane 1. d, efficiency of competition of unlabeled −2554T and −2554G oligomers for NF-κB family of transcription factors. The nuclear proteins in PHA blasts that bound a radiolabeled consensus NF-κB-binding oligomer were competed with excess unlabeled NF-κB oligomers (lanes 2–5), −2554G (lanes 6–9), or −2554T (lanes 10–13) oligomers. An equimolar amount of unlabeled −2554T oligomer was much more efficient in competing NF-κB/Rel family of proteins than the −2554G oligomer (compare lanes 6–9 versus 10–13). However, compared with the unlabeled consensus NF-κB oligomer (lanes 2–5), both the −2554G (lanes 6–9) and −2554T (lanes 10–13) oligomers were significantly less efficient in competing for NF-κB proteins.

Direct evidence that −2554T and −2554G bound to the NF-κB family of transcription factor and two additional novel factors (NF6 and NF7) with differing avidity (−2554T > −2554G) are shown in Fig. 5b (compare intensity of bands in lanes 1 and 2) and Fig. 5 c (compare intensity of bands in lanes 1 and 10). Indirect evidence of the greater avidity of −2554T for NF-κB binding is further supported by the differing efficiency of the unlabeled −2554G and −2554T oligomers to compete for the NF-κB proteins bound to the radiolabeled NF-κB oligomer (−2554T > −2554G; Fig. 5d). However, both the unlabeled −2554G and −2554T oligomers were significantly less efficient than the unlabeled NF-κB oligomer in competing for the nuclear proteins bound to the radiolabeled NF-κB oligomer (Fig. 5d).

Collectively, based on these EMSA and supershift experiments, as well as the densitometric analysis, the affinity of various members of the NF-κB family of transcription proteins for the three oligomers tested, i.e. CCR5 −2554G, CCR5 −2554T, and the consensus NF-κB oligomer, varied and could be expressed as −2554G ≪ −2554T ≪ NF-κB. However, in addition to the NF-κB proteins, the −2554G and −2554T oligomers bound to two nuclear factors (NF6 and NF7) that were not bound by the consensus NF-κB oligomer, and the affinity of NF6 and NF7 for the three oligomers was NF6 < −2554G ≪ −2554T. We observed some donor-to-donor variability in the avidity with which the NF4-NF7 complexes bound to the −2554G/G/T oligomers (data not shown). However, in each case the aforementioned relationship of NF4 and NF5 (−2554G ≪ −2554T ≪ NF-κB) and NF6 and NF7 (NF-κB ≪ −2554G ≪ −2554T) for the three oligomers tested remained constant.

Functional Effects of Variation in Homologous CCR5 cis-Regulatory Regions of Human and Non-human Primates: Differential Transcriptional Activity and Altered DNA/Nuclear Factor Binding—The region encompassing human CCR5 −2761 and −1934 confers strong promoter activity in different cellular environments (20, 36–38). To determine whether the homologous genomic region in AGM conferred similar or dif-
different promoter activities, we tested the promoter strengths of various human and AGM (sabaeus) constructs in HEK, K562, and COS cell lines (Fig. 6a). Constructs that originated at −2761, −2570, and −2275 had the highest transcriptional efficiency (Fig. 6a). Relative to human construct H2, the homologous cis-regulatory region in sabaeus (S2) had higher promoter activity in all three cellular environments tested, and S1 and S3 had higher promoter activity than H1 and H3, respectively, in COS cells (Fig. 6a).

Sequence differences in the regulatory regions of human and non-human primate CCR5 could potentially lead to variation in the binding of nuclear factors to homologous cis-regions. We examined the avidity with which NF-κB transcription factors bound to a region in OWM CCR5 that is homologous to that spanning human CCR5 −2554. The NF-κB-binding site in humans and great apes extends from −2558 to −2549 (Fig. 5a). It is striking that at the position homologous to human CCR5 −2554, in addition to several human haplogroups, great apes, OWM, and NWM all contain a −2554G. However, at the position homologous to human CCR5 −2550C in OWMs there is a C (cytosine) → T (thymidine) change (Fig. 5a), and this is predicted to disrupt the consensus sequence for binding of the p65 subunit (39). A similar change is observed in NWMs. An oligomer derived from a region in OWMs that is homologous to human CCR5 −2558 to −2549 in humans/great apes (Table 1) was tested for its ability to bind to the NF-κB/c-Rel transcription factors. For these experiments, PBMCs were obtained from a sabaeus AGM and a vervet AGM that were not infected with SIV. These PBMCs were stimulated with PHA and IL-2, and nuclear extracts were prepared from the activated cells. In parallel control experiments, human PBMCs were also stimulated in an identical manner and nuclear extracts prepared. In EMSAs, a consensus NF-κB-binding oligomer bound to a complex of factors in OWM nuclear extracts from PHA blasts and by supershift assays appeared to contain p65 (moderate supershift), p50 (strong supershift), and c-Rel (weak supershift) but not p52 (Fig. 6b). In contrast, the radiolabeled OWM CCR5 −2554G oligomer (Table 1) did not bind specifically to any nuclear proteins in PHA blasts, and this oligomer and the NWM oligomers competed inefficiently for the nuclear factors that bound to a radiolabeled NF-κB consensus oligomer (Fig. 6c).

Similar supershift results were obtained regardless of whether the nuclear extracts were derived from sabaeus/vervet or human PHA blasts, suggesting that the NF-κB proteins are conserved in humans and non-human primates.

**DISCUSSION**

Role of Comparative Genomics in Understanding the Pathogenesis of HIV/SIV—There has been substantial effort to understand the evolution of HIV and SIV. However, there is little information about the evolution or even interspecies variation of the host determinants of HIV-1 and SIV pathogenicity. CCR5 is one of the major determinants of cell entry for HIV-1 and SIV-1 (10, 11), and thus expression levels of CCR5 are likely to play an important role in virus transmission and disease pathogenesis. We and others (14–17, 40–45, 67) have demonstrated that genetic variation in the CCR5 cis-regulatory regions is much higher than rare mutations like the 32-bp deletion of the ORF (16). Undoubtedly, these rare variants are useful for understanding further the pathogenesis of HIV-1 infection; however, the common promoter variants will likely to be of greater importance than the rare variants, especially if they are associated with phenotypic consequences.

To initiate studies to define mechanisms by which genetic variation in human and non-human primate CCR5 might contribute to the differential pathogenesis of these two closely related lentiviruses, and influence the inter-individual and inter-population variation in susceptibility to HIV-1 infection, in this study we (i) developed a clear understanding of the extent, organization, and evolution of genetic variation in the coding and cis-regulatory regions of human and non-human primate CCR5; (ii) developed an extensive data base of genomic sequences to create appropriate intra- and interspecies haplotype-specific promoter constructs that we show are associated with differential intra- and interspecies haplotype-specific transcriptional activity; (iii) identified common intra- and interspecies cis-regulatory polymorphisms associated with altered DNA/protein interactions, including intra- and interspecies differential binding of various members of the NF-κB family of transcription factors; and (iv) show that because of mutations at intron-exon splice donor-acceptor sites, CCR5 mRNA processing in humans and non-human primates is not
identical. Collectively, these findings suggest that the generation of multiple CCR5 haplotypes and mRNA isoform diversity might be a general evolutionary strategy of primates to pathogens such as HIV/SIV. The intra- and interspecies differences in CCR5 haplotypes and mRNA isoforms, haplotype-specific transcriptional activity, and DNA/nuclear factor-binding might explain, in part, the differential pathogenesis of HIV and SIV and potentially the different responses to HIV among human populations.

**Evolution of Primate CCR5 ORF and cis-Regulatory Regions**—The ORF and the cis-regulatory region of human CCR5 exhibited higher nucleotide variability than the reported average values (46), and variation in human CCR5 was clearly higher than has been commonly appreciated (14, 17). Moreover, the ascertainment bias introduced by our initial sampling of individuals homozygous for different CCR5 SNPs (see “Experimental Procedures” for details) suggested that our estimate of sequence diversity is conservative. Despite similar levels of genetic diversity between primate groups, analysis of the genetic distances (affinities) among different groups of primates revealed an interesting pattern. The genetic distance between OWM and hominoids estimated from the CCR5 cis-regulatory region was approximately four times larger than the distance estimated from the CCR5 ORF. This may be the consequence of relaxed selection on a noncoding region of CCR5 versus the CCR5 ORF. This pattern could also be due to the effects of different selective forces and/or selection for different polymorphisms in the CCR5 cis-regulatory region among OWM and hominoids. If the former is true, estimates of the genetic affinities among primate groups from the cis-regulatory region of CCR5 and the CCR5 ORF should be comparable. Only the genetic distance between OWM and hominoids should be different if the latter is true. Thus, it is conceivable that differences in the effects of natural selection may be responsible, in part, for the divergence observed in the cis-regulatory region of hominoid CCR5. Additionally, other forces that shape genetic evolution (e.g., genetic drift, population history) may have differed among primates. Nevertheless, the accumulation of different polymorphisms may have affected the transcriptional/translational activity of CCR5 permitting OWM and hominoids to modulate responses to different repertoires of pathogens.

Many of the non synonymous substitutions in the CCR5 ORF were clustered in the region encoding the NH2 terminus of CCR5. HIV-1 appears to interact via gp120 with the ligand-specific binding site of CCR5, and the NH2 terminus of CCR5 determines, in part, the specificity of this binding (47). A DNdS of >1.0 for the NH2 terminus of CCR5 suggested that positive selection may have had an important role in generating variation in this region of the CCR5 ORF in hominoids. Thus, certain amino acid substitutions in the NH2 terminus of CCR5 may represent selection of variant phenotypes (and, hence genotypes) following interaction of our hominoid ancestors with members of the lentivirus family. More importantly, these results indicate that the NH2 terminus of CCR5 may be a preferred target for interventions to prevent HIV-1 entry into human macrophages (48). Nevertheless, the bulk of polymorphisms in CCR5 were found in the cis-regulatory regions. Thus, it is important that this variation be organized in such a manner as to be useful for understanding the effect of these polymorphisms on the pathogenesis of HIV-1. This was the logic behind organizing CCR5 cis-regulatory region haplotypes into a rooted phylogenetic network.

**Human CCR5 Haplotype Classification and Differential Haplotype-specific Promoter Activity**—A major limitation of previous attempts to understand the organization of human CCR5 haplotypes has been a lack of an appropriate outgroup to root the ancestral CCR5 haplotype (15). Here, we established the ancestral CCR5 haplotype, and we used this information to create a framework for a biologically based classification and nomenclature of human CCR5 haplotypes. The organization of the complex patterns of CCR5 polymorphisms into evolutionarily meaningful relationships has at least three merits.

First, it provides a framework for understanding the association between different CCR5 haplotypes and HIV-1 disease progression or transmission. Because of extensive sequence variation, comprehensive genotyping of each individual for every CCR5 polymorphism would be costly, labor-intensive, and inefficient. In contrast, a phylogenetic network of CCR5 haplotypes forms a basis for grouping CCR5 haplotypes whose relationships to each other can be defined unambiguously by a single or few polymorphisms. This forms the rationale for grouping CCR5 haplotypes that are closely related to each other (e.g., all descendants of a shared ancestral haplotype). For example, all haplotypes that are characterized by the −2733G polymorphism but lack the CCR5 Δ32 mutation can be grouped into HHG*1, and CCR5 −1835T-bearing haplotypes lacking CCR2−641 are classified as HFF*1. By extensive genotyping of human subjects, we found the prevalence of HFF*1 haplotypes in worldwide populations to vary from −1 to 12% (16).

The phylogenetic network of CCR5 haplotypes also potentially helps to lessen haplogroup/haplotype misclassification and will facilitate genotype-phenotype analyses. For example, McDermott et al. (17), reported recently that an allele (59029A allele) with genotypic features similar to what can be considered as a −2459A allele was associated with higher transcriptional efficiency and that homozygosity for this allele was associated with accelerated disease progression. Similarly, homozygosity for an allele designated as the P1 allele was also shown to be associated with disease progression (14). However, our data demonstrate that these two alleles (P1 or 59029A) are a mixture of at least three haplogroups that share −2459A and −2135C (HHE, HFF*1, and HHG*1). Based on sequence data from −2554 to −1951, Martin et al. (14) described nine additional CCR5 alleles designated as P2–P10, whereas we show that this represents a subset of the total CCR5 haplotype pool. Our data also demonstrate that P2, P3, and P4 represent alleles that correspond to alleles within HHAA, HHBD, and HHC, respectively, and that alleles labeled as P5–P10 are likely to correspond to alleles within HHAA, HHC, or HHBD. Thus, the organization of CCR5 haplotypes into an evolutionary framework can potentially minimize the confounding that might occur by mixing SNPs and/or haplotypes with different evolutionary histories and phenotypic effects.

Second, this classification will enable studies aimed at exploring further the basis for the varied distribution of CCR5 haplotypes among contemporary human populations. For example, the allele frequency of the ancestral CCR5 haplotype (HHA) is higher in individuals of African descent (>0.20) than in Caucasians (−0.09) and peaks in African Mbuti Pygmies (0.71), in whom the prevalence of HIV-1 infection appears to be very low (16). Although we found no evidence that HHA affords resistance to infection, this haplotype was associated with HIV-1 disease retardation in African Americans but not in Caucasians. Determining the biological basis for the varied frequencies of CCR5 haplotypes among populations will be important in evaluating differences in susceptibility and disease progression among these groups. It should be noted that the phylogenetic network presented in this study is a relatively robust and objective depiction of the relationships among the polymorphisms in an important cis-regulatory region of CCR5. As more sequence data are incorporated, the topology of some of the branches is likely to change, and these changes can be
easily incorporated into this network. For example, further sequencing of regions 5’ and 3’ of the CCR5 haplotypes on the branch tips of this network has already identified additional polymorphisms (data not shown) that further distinguish haplotypes from each other. Consequently, we think it is premature to introduce any further nomenclature (i.e. beyond haplogroups) until the sequence 5’ and 3’ to the CCR5 haplotypes shown in this paper are better characterized. Otherwise, the nomenclature of CCR5 haplotypes will be changing frequently as additional polymorphisms are described and potentially lead to greater confusion among investigators.

The distribution and placement of CCR5 polymorphisms relative to one another in the network of haplotypes will also facilitate investigation of the evolutionary forces that have driven these haplotypes to varying frequencies among different human populations. For example, population-specific deleterious or protective mutations (e.g. Δ32) that were found near the tips of branches may have arisen more recently than polymorphisms embedded deeper in the network. This suggests the following: 1) SNPs at −2733, −2554, −2459, −2135, and −1835 are older than the Δ32, −2132T, −2086G, and CCR2-V64I polymorphisms; 2) CCR2-V64I predates the Δ32 mutation; and 3) contrary to previous assertions, it is more likely that the ancestral state of the −2459 residue is guanine and not adenine (17).

The third advantage of organizing CCR5 variation into a phylogenetic network is that it increases the efficiency of identifying specific sequence motifs in the cis-regulatory region of CCR5 that might produce different effects in vitro. We demonstrated that one of the mechanisms underlying the effects of different CCR5 haplotypes might be differential transcriptional efficiency. Yet, it would be more difficult to interpret these findings if the ancestor-descendant relationships between polymorphisms were unknown. Promoter analysis of constructs spanning the major SNPs that distinguish CCR5 haplogroups demonstrate that nucleotide substitutions in the cis-regulatory regions of CCR5 produce differences in transcriptional activity. For example, in K562 cells (Fig. 4a) and Jurkat cells (data not shown), the ancestral HHA haplotype-promoter construct consistently demonstrated the lowest transcriptional activity, whereas the transcriptional activity of HHF haplotypes was the highest of the haplogroup-specific constructs tested. Our analysis of the association of CCR5 haplogroups and HIV-1 disease progression suggests that HHA and HHF∗2 haplotypes are both associated with HIV-1 disease retardation in adults (16). This suggests that correlating in vitro findings of differences in haplotype-specific transcriptional efficiencies to differences in the surface expression of CCR5 and/or the disease-modifying effects of CCR5 haplotypes is going to be a difficult endeavor.

Implications of Altered DNA/Nuclear Factor Interactions Due to Primate CCR5 cis-Regulatory Region SNPs—There are several lines of evidence from other gene systems to indicate that genetic diversity is an important determinant of disease susceptibility. For example, a SNP that creates an Oct-1-binding site and HIV-1 disease progression suggests that HHA and HHF∗2 haplotypes are both associated with HIV-1 disease retardation in adults (16). This suggests that correlating in vitro findings of differences in haplotype-specific transcriptional efficiencies to differences in the surface expression of CCR5 and/or the disease-modifying effects of CCR5 haplotypes is going to be a difficult endeavor.

Implications of Primates CCR5 mRNA Diversity—As in other gene systems (55–58), the 5’ untranslated regions of CCR5 could be an important determinant of expression through an effect on either RNA turnover or translational efficiency. In some instances, sequences in the 5’-UTRs of mRNA species contribute to secondary structure and binding of associated proteins that might regulate RNA stability (59). However, most studies on the effects of the 5’-UTR have found differences in translational efficiency. Secondary structure, decreased accessibility to the m7G cap, and some sequence elements in the 5’-UTR all decrease the efficiency with which mRNA species produce protein (55, 60, 61). Moderately stable hairpin structures close to the 5’ end are thought to block access to the m7G cap, whereas structures located downstream apparently inhibit translation by decreasing processivity of the ribosomal complex. Hence, it is possible that specific sequences and/or structures in the 5’-UTRs of primates CCR5 could influence their translation.

In this context, we identified novel human CCR5 mRNA sequences and demonstrated that these sequences and the complex RNA structure of human CCR5 are conserved in OWMs and apes. Of note, modifications of primate CCR5 DNA by mutations alter the CCR5 mRNA isoforms produced, thus
increasing the diversity of phenotypes. Also, the predicted hairpin structures of the primate mRNAs were strikingly different (data not shown). These findings support the hypothesis that both different CCR5 mRNA isoforms and polymorphisms in the distinct 5'-UTRs that compose these RNA species might influence CCR5 cell-surface expression by regulating gene expression at a post-transcriptional level. Alternatively, distinct secondary structures such as stem loops in the different CCR5 RNA isoforms could potentially influence translation efficiency. These hypotheses and the role of different human CCR5 mRNA transcripts in HIV pathogenesis and the effects of the variable presence or absence of these transcripts among non-human primates on SIV pathogenesis are currently under investigation.

SIVs in their natural host African primate have most likely arisen through coevolution with their respective host, suggesting a long period of adaptive evolution (62, 63). For sooty mangabeys and AGMs including sabaeus monkeys, a lack of pathogenicity has been associated with an overall lower viral burden in peripheral blood cells as compared with HIV-infected humans (64). However, high viral viremias are maintained in these monkeys despite significantly fewer infected cells, suggesting fundamental differences in host-virus dynamics. In part, these differences may result from subtle differences in the levels of expression of coreceptors, including CCR5. Although CCR5 appears to be the main coreceptor used by a variety of SIVs and HIV, other coreceptor usage could also be modulated in the natural host (65, 66). Our data would support the notion that CCR5 haplotype and mRNA isoform diversity, differences in transcriptional activity, and altered DNA/nuclear factor binding may collectively result in subtle differences in CCR5 expression levels, and possibly tissue tropism for SIVs, leading to overall fewer infected cells and hence a non-pathogenic state. Furthermore, our data also emphasize an important role for generating and maintaining polymorphisms in the regulatory regions and 5'-UTR of CCR5. These polymorphisms and the trans-acting nuclear factors that bind them are likely to be important determinants in HIV and SIV pathogenesis. Studies directed at understanding the patterns of CCR5 expression in the context of host progression may provide clues to the disparate nature of host viral relationships.

Acknowledgments—We thank Drs. R. A. Clark and L. Jorde for helpful comments and strong support; N. Nichols, Leona Mosher, Ai-Hsin Cheng, and LeeAnn Lam for technical assistance, Dr. R. Dhanda for statistical analysis; Dr. Joe P. Flanagan and the Houston Zoo for providing non-human primate blood samples; and A. S. Ahuja for forbearance. We also thank the reviewers of this manuscript for their insightful comments and suggestions. S. S. A. and S. K. A. are deeply indebted to the Directors of the Robert J. Kleberg Jr., and Helen C. Kleberg Foundation for their strong support. We apologize for our inability to reference several additional excellent articles on this subject because of space constraints.

Note Added in Proof—While this manuscript was in review, by examining a CCR5 cis-regulatory region that spans −2459 to −1548 of humans, chimpanzee, pig-tailed macaque, and sooty mangabey, Tang et al. (67) described an ancestral CCR5 allele designated as CCR5P*0102 allele. This allele is a combination of HHA and HHB. The CCR5P*0201 allele described by Tang et al. contains both HHE and HHG.

REFERENCES


6 J. S. Allan, unpublished data.
Evolution of Human and Non-human Primate CC Chemokine Receptor 5 Gene and mRNA: POTENTIAL ROLES FOR HAPLOTYPE AND mRNA DIVERSITY, DIFFERENTIAL HAPLOTYPE-SPECIFIC TRANSCRIPTIONAL ACTIVITY, AND ALTERED TRANSCRIPTION FACTOR BINDING TO POLYMORPHIC NUCLEOTIDES IN THE PATHOGENESIS OF HIV-1 AND SIMIAN IMMUNODEFICIENCY VIRUS


doi: 10.1074/jbc.M000169200 originally published online March 9, 2000

Access the most updated version of this article at doi: 10.1074/jbc.M000169200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

Supplemental material:
http://www.jbc.org/content/suppl/2000/07/10/275.25.18946.DC1

This article cites 66 references, 17 of which can be accessed free at http://www.jbc.org/content/275/25/18946.full.html#ref-list-1