Am. J. Hum. Genet. 77:869–886, 2005

The Heritage of Pathogen Pressures and Ancient Demography in the Human Innate-Immunity *CD209/CD209L* Region

Luis B. Barreiro,[1,2] Etienne Patin,[1] Olivier Neyrolles,[2] Howard M. Cann,[3] Brigitte Gicquel,[2] and Lluís Quintana-Murci[1]

[1]Centre National de la Recherche Scientifique FRE 2849, Unit of Molecular Prevention and Therapy of Human Diseases, and [2]Unité de Génétique Mycobactérienne, Institut Pasteur, and [3]Fondation Jean Dausset, Centre d'Etudes du Polymorphisme Humain (CEPH), Paris

The innate immunity system constitutes the first line of host defense against pathogens. Two closely related innate immunity genes, *CD209* and *CD209L*, are particularly interesting because they directly recognize a plethora of pathogens, including bacteria, viruses, and parasites. Both genes, which result from an ancient duplication, possess a neck region, made up of seven repeats of 23 amino acids each, known to play a major role in the pathogen-binding properties of these proteins. To explore the extent to which pathogens have exerted selective pressures on these innate immunity genes, we resequenced them in a group of samples from sub-Saharan Africa, Europe, and East Asia. Moreover, variation in the number of repeats of the neck region was defined in the entire Human Genome Diversity Panel for both genes. Our results, which are based on diversity levels, neutrality tests, population genetic distances, and neck-region length variation, provide genetic evidence that *CD209* has been under a strong selective constraint that prevents accumulation of any amino acid changes, whereas *CD209L* variability has most likely been shaped by the action of balancing selection in non-African populations. In addition, our data point to the neck region as the functional target of such selective pressures: *CD209* presents a constant size in the neck region populationwide, whereas *CD209L* presents an excess of length variation, particularly in non-African populations. An additional interesting observation came from the coalescent-based *CD209* gene tree, whose binary topology and time depth (~2.8 million years ago) are compatible with an ancestral population structure in Africa. Altogether, our study has revealed that even a short segment of the human genome can uncover an extraordinarily complex evolutionary history, including different pathogen pressures on host genes as well as traces of admixture among archaic hominid populations.

Introduction

Infectious diseases have been paramount among the threats to health and survival for most of human evolutionary history (Haldane 1949; Lederberg 1999; Harpending and Rogers 2000; Cooke and Hill 2001). The interaction of the human host with a wide variety of pathogens has been accompanied by genetic adaptations to spatially and temporally fluctuating selective pressures imposed by the infectious agents. Numerous studies have sought the genetic imprint of natural selection imposed by pathogen pressures in human genes involved in immune response or, more generally, in host-pathogen interactions (Vallender and Lahn 2004). For example, natural selection has acted on such genes as *MHC*, *β-globin*, *G6PD*, *IL-2*, *IL-4*, *TNFSF5*, the Duffy blood group genes, and *CCR5* (Ohta 1991; Hughes et al. 1994; Flint et al. 1998; Hamblin and Di Rienzo 2000; Tishkoff et al. 2001; Bamshad et al. 2002; Sabeti et al. 2002; Verrelli et al. 2002). However, little is known about genetic variation of genes involved in direct recognition of pathogens, or pathogens' products, and virtually no studies have investigated the extent to which pathogens have exerted selective pressures on the innate immune system.

The phylogenetically ancient innate immune system governs the initial detection of pathogens and stimulates the first line of host defense (Medzhitov and Janeway 1998*a*, 2000, 2002; Janeway and Medzhitov 2002). Recognition of pathogens is mediated by phagocytic cells through germline-encoded receptors, known as "pattern recognition receptors," which detect pathogen-associated molecular patterns that are characteristic products of microbial physiology (Kimbrell and Beutler 2001; Janeway and Medzhitov 2002). This initial interaction is then translated into a set of endogenous signals that ultimately lead to the induction of the adaptive immune response (Medzhitov and Janeway 1998*b*).

In recent years, the C-type lectin receptors have re-

Received July 19, 2005; accepted for publication August 26, 2005; electronically published September 29, 2005.

Address for correspondence and reprints: Dr. Lluís Quintana-Murci, Centre National de la Recherche Scientifique FRE 2849, Unit of Molecular Prevention and Therapy of Human Diseases, Institut Pasteur, 25, rue Dr. Roux, 75724 Paris Cedex 15, France. E-mail: quintana@pasteur.fr

0002-9297/2005/7705-0015$15.00

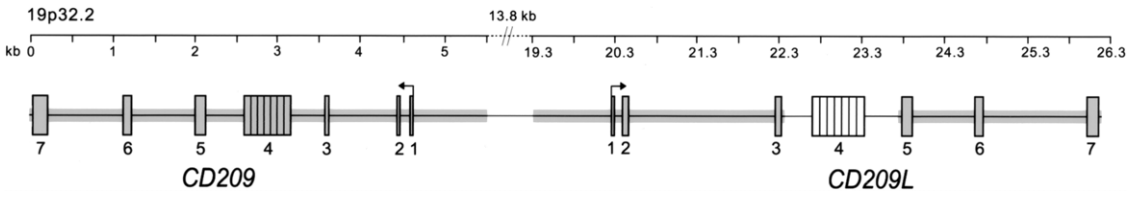

Figure 1 Scaled diagram of the *CD209/CD209L* genomic region. Sequenced regions are represented in gray. For *CD209,* we sequenced a total of 5,500 bp per chromosome, and, for *CD209L,* 5,391 bp per chromosome. The neck region corresponding to exon 4 and composed of seven coding repeats is also shown.

ceived much attention in the area of innate immunology, the results of which were novel functional insights into the primary interface between host and pathogens (Medzhitov 2001; Cook et al. 2003; Fujita et al. 2004; Geijtenbeek et al. 2004; McGreal et al. 2004). In this context, two prototypic members of the C-type lectin–receptor family are particularly interesting, since they can act as both cell-adhesion receptors and pathogen-recognition receptors. These lectins include CD209 (DC-SIGN: dendritic cell–specific ICAM-3 grabbing nonintegrin [MIM 604672]) and its close relative CD209L (L-SIGN: liver/lymph node–specific ICAM-3 grabbing nonintegrin [MIM 605872]) (Curtis et al. 1992; Geijtenbeek et al. 2000*b,* 2004; Soilleux et al. 2000; Pohlmann et al. 2001). These lectin-coding genes are located on chromosome 19p13.2-3, within an ~26-kb segment, and result from a duplication of an ancestral gene (Bashirova et al. 2003; Soilleux 2003). An additional characteristic of both *CD209* and *CD209L* is the presence of a neck region, primarily made up of seven highly conserved 23-aa repeats, that separates the carbohydrate-recognition domain involved in pathogen binding from the transmembrane region. This neck region presents high nucleotide identity between repeats, both within each molecule and between *CD209* and *CD209L.* It has been shown that this region plays a crucial role in the oligomerization and support of the carbohydrate-recognition domain; therefore, it influences the pathogen-binding properties of these two receptors (Soilleux et al. 2000, 2003; Feinberg et al. 2005). In regard to expression profiles, *CD209* is expressed primarily on phagocytic cells, such as dendritic cells and macrophages, whereas *CD209L* expression is restricted to endothelial cells in liver and lymph nodes (Bashirova et al. 2001; Soilleux et al. 2001, 2002). As pathogen-recognition receptors, the two lectins have been shown to recognize a vast range of microbes, some of which are of major public health importance (Geijtenbeek et al. 2004). Indeed, CD209 captures bacteria such as *Mycobacterium tuberculosis, Helicobacter pylori,* and certain *Klebsiela pneumonia* strains; viruses such as HIV-1, Ebola virus, cytomegalovirus, hepatitis C virus, Dengue virus, and SARS-coronavirus; and parasites like *Leishmania pifanoi* and *Schistosoma mansoni* (Geijtenbeek et al. 2000*a,* 2003; Alvarez et al. 2002; Colmenares et al. 2002; Halary et al. 2002; Appelmelk et al. 2003; Lozach et al. 2003; Tailleux et al. 2003; Tassaneetrithep et al. 2003; Bergman et al. 2004; Marzi et al. 2004). With regard to CD209L, studies to date have shown an interaction with a variety of viruses, including HIV, hepatitis C, Ebola, and coronavirus, as well as with the parasite *Schistosoma mansoni* (Bashirova et al. 2001; Alvarez et al. 2002; Gardner et al. 2003; Jeffers et al. 2004; Van Liempt et al. 2004). In this context, the efficiency of the two lectins in pathogen recognition and subsequent processing may have important consequences for the quality of host immune responses and consequent pathogen control and/or clearance.

An important step forward in the understanding of human adaptation to pathogens and control of infectious diseases includes the description of quality and quantity of genetic variation in genes involved in host recognition of infectious agents. Given the direct interaction of CD209 and CD209L with a large variety of pathogens, the *CD209/CD209L* genomic region provides an excellent model system to illustrate the extent to which pathogens have exerted selective pressures on host immunity genes. An additional feature that makes these genes highly interesting in evolutionary studies is that they are likely to have been influenced by similar genomic forces (recombination, mutation rates, etc.) because of their close physical proximity (~15 kb), high nucleotide (73%) and amino acid (77%) identity, and identical exon-intron organization (Soilleux 2003) (fig. 1). In addition, it has been proposed that gene duplication of immunity genes is a molecular strategy developed by the host to enlarge its defense potential (Ohno 1970; Trowsdale and Parham 2004). A number of immune-system gene families have evolved, by gene duplication followed by natural selection, to provide responses to a wider range of pathogens, with well-documented examples in immunoglobulin and *MHC* genes (Hughes et al. 1994; Ota et al. 2000). In this context, duplicated genes in *cis,* like *CD209* and *CD209L,* may have undergone differential selective pressures to enlarge the defense role of these lectins. To address these

complex issues, we performed a sequence-based survey of the entire *CD209/CD209L* region in a panel of individuals of different ethnic origins. Here, we report evidence showing that these two closely related innate immunity genes have gone through completely different evolutionary processes that are reflected in their current patterns of diversity. In addition, our study provides novel insights into how pathogens have shaped the patterns of variability of immunity genes resulting from gene duplication.

Material and Methods

Population Samples

Sequence variation of the *CD209/CD209L* region was determined in 41 sub-Saharan Africans, 43 Europeans, and 43 East Asians, in a total of 254 chromosomes from the Human Genome Diversity Panel (HGDP)–CEPH panel (Cann et al. 2002). More-detailed information about the composition of the three major ethnic groups can be found in table 1. The variation in the repeat number of the neck region of *CD209* and *CD209L* was defined in the entire HGDP-CEPH panel, comprising 1,064 DNA samples from 52 worldwide populations. In addition, the orthologous regions for both genes were sequenced in four chimpanzees (*Pan troglodytes*).

Molecular Analyses

The sequenced fragments of the *CD209/CD209L* genomic region are shown in figure 1. The entire *CD209* region—including exons, introns, and ~1 kb of the 5′ UTR corresponding to the promoter region—was sequenced, for a total of 5.5 kb per individual. For *CD209L,* we sequenced a total of ~5.4 kb per individual, following the same approach used for *CD209,* with the exception of the neck region. That region was genotyped for its number of repeats, since it turned out to be highly polymorphic, which prevented the sequencing process. Genotyping was performed by a single PCR amplification followed by migration in 2% agarose gels. Human primers were used to both amplify and sequence the orthologous regions in chimpanzees. However, because of polymorphisms specific to the chimpanzee lineage, we could not obtain the entirety of the sequence. Thus, 4.9 kb (90% of the total) of the chimpanzee *CD209* sequence were obtained, and 5.3 kb (98% of the total) of *CD209L.* Detailed information on primer sequences and PCR amplification conditions is available on request. All nucleotide sequences were obtained using the Big Dye terminator kit and the 3100 automated sequencer from Applied Biosystems. Sequence files and chromatograms were inspected using the GENALYS software (Takahashi et al. 2003; Centre National de Genotypage). As a measure of quality control, when new mutations were identified in primer binding regions, new primers were designed and sequence reactions were repeated, to avoid allele-specific amplification. All singletons observed in our data set were systematically reamplified and resequenced.

Table 1

Individual Composition of the Study Populations for *CD209* and *CD209L* Sequence Diversity

The table is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Statistical Analyses

On the basis of the levels of diversity observed in the *CD209/CD209L* genomic region, we calculated the average number of pairwise differences (π) and the Watterson's estimator (θ_w) (Watterson 1975). Under the standard neutral model of a randomly mating population of constant size, these are unbiased estimators of the population mutation rate $\theta = 4N_e\mu$, where N_e is the diploid effective population size and μ is the mutation rate per generation per site. To test whether the frequency spectrum of mutations conformed to the expectations of this standard neutral model, we calculated Tajima's D (Tajima 1989) and Fay and Wu's H tests (Fay and Wu 2000). P values for the different tests were estimated from 10^4 coalescent simulations under an infinite-site model, with use of a fixed number of segregating sites and the assumption of no recombination, which has been shown to be a conservative assumption (Gilad et al. 2002). In parallel, we estimated P values for all these tests, using the empirical distribution obtained from sequencing data of 132 genes in a panel of 24 African Americans and 23 European Americans (Akey et al. 2004). All these analyses, together with the interspecies McDonald-Kreitman (McDonald and Kreitman 1991) and K_A/K_S (Kimura 1968) tests, were performed using the DnaSP package (Rozas et al. 2003). Genetic distances between populations (F_{ST}) and heterozygosity values were estimated using the Arlequin package (Schneider et al. 2000). F_{ST} statistical significance was assessed using 10,000 bootstrap replications. To bear out a deficit or an excess of heterozygosity in the neck region of *CD209* and *CD209L,* we used BOTTLENECK (Cornuet and Luikart 1996) to compute for each geographic region, the distribution of the heterozygosity expected from the observed number of alleles, given the sample size (n) under the assumption of mutational-drift equilibrium. This distribution was obtained through simulation of the coalescent process of n genes under two mutational models, the infinite-site model and the stepwise mutation model. In addition, to obtain information on the fraction of genetic variance in the neck region that is due to intra- and interpopulation differences, we performed an anal-

ysis of molecular variance (AMOVA), using the Arlequin package (Schneider et al. 2000). The AMOVA results were compared with those of 377 microsatellites analyzed in the same population panel (Rosenberg et al. 2002).

Haplotype reconstruction was performed by use of the Bayesian statistical method implemented in Phase (v.2.1.1) (Stephens and Donnelly 2003). We applied the algorithm five times, using different randomly generated seeds, and consistent results were obtained across runs. After haplotype reconstruction, linkage disequilibrium (LD) between pairs of SNPs was computed using Lewontin's D' index (Lewontin 1964). For this analysis, only markers presenting a minimum allele frequency (MAF) of 10% were considered, since rare alleles have been shown to present a higher probability of being in significant LD than do common ones (Reich et al. 2001). The graphic display of the LD plots was constructed using GOLD (Abecasis and Cookson 2000; Center for Statistical Genetics). To support the existence of a recombination hotspot in the region under study, we used the hotspot-recombination model implemented in Phase (v.2.1.1). Under this model, we assumed that there was, at most, one hotspot of unknown position. We then estimated the background population-recombination rate (ρ) and the relative intensity of any recombination hotspot. To obtain better estimates, we increased 10 times the number of iterations of the final run of the algorithm. All our estimations were obtained by averaging results of five independent runs with use of different seed numbers. Since the model used is Bayesian, we could also estimate, for each population, the posterior probability of a hotspot of intensity >1 ($\lambda > 1$) and >10 ($\lambda > 10$).

We obtained the gene tree and estimated the time of the most recent common ancestor (T_{MRCA}) for *CD209,* using the maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994). The mutation rate μ for each gene was estimated on the basis of the net divergence between humans and chimpanzees and under the assumption both that the species separation occurred 5 million years ago (MYA) and of a generation time of 20 years. Using this μ and θ maximum likelihood (θ_{ML}), we estimated the effective population size parameter (N_e). With the assumption of a generation time of 20 years and the estimated N_e, the coalescence time, scaled in $2N_e$ units, was converted into years. The coalescence process implemented in SIMCOAL2 (Laval and Excoffier 2004) allowed us to estimate the probability of the T_{MRCA} for *CD209,* through 2×10^4 simulations, with use of both the number of observed segregating sites and the estimated N_e.

Results

We determined sequence diversity in the *CD209* and *CD209L* genes (fig. 1) as well as length variation of the neck region in 254 chromosomes originating from three major ethnic groups: sub-Saharan Africans, Europeans, and East Asians. In addition, the orthologous sequences were obtained in four chimpanzees, to infer the ancestral state at each site, to estimate the divergence between humans and chimpanzees, and to perform a number of interspecies neutrality tests.

Patterns of Nucleotide and Haplotype Diversity in the CD209/CD209L *Region*

For *CD209,* we identified a total of 79 SNPs and 2 indels, including 5 nonsynonymous, 5 synonymous, and 71 noncoding variants. The five nonsynonymous SNPs were all located in the neck region (exon 4): SNPs 1839 (Arg→Gln), 1888 (Glu→Asp), and 1908 (Arg→Gln) achieved a frequency of ~15%, and SNP 1970 (Leu→Val), a frequency of 6%. These mutations were restricted to the African sample. SNP 1472 (Ala→Thr) was observed as a singleton in an East-Asian individual. For *CD209L,* we identified 64 SNPs and 2 indels, including 4 nonsynonymous and 62 noncoding variants. The four nonsynonymous variants were located in different exons: SNP 141 (Thr→Ala) in exon 2, SNP 3476 (Asp→Asn) in exon 5, SNP 4268 (Thr→Ala) in exon 6, and SNP 5580 (Arg→Gln) in exon 7. All these mutations were singletons except SNP 3476, which presented high frequencies for its derived allele in all geographic regions: 97.6% in Africans, 57% in Europeans, and 77% in East Asians. All variable sites were in Hardy-Weinberg equilibrium for both *CD209* and *CD209L,* after Bonferroni correction for multiple testing.

The allelic composition of *CD209* and *CD209L* haplotypes and their frequency distribution in the three major ethnic groups is illustrated in figure 2, along with the haplotype composed of the ancestral allelic state of each SNP inferred from chimpanzee data. For *CD209,* we identified 42 different haplotypes, with an overall heterozygosity of 84% (table 2). Three major haplotypes (H2, H29, and H40) accounted for ~50% of the African variability, whereas they were at very low frequency (H2 at ~5%) or absent (H29 and H40) in Europeans and East Asians (fig. 2*A*). In turn, the two haplotypes (H1 and H3) that accounted for 58% and 83% of the European and East Asian variability, respectively, were observed at very low frequency (H1 at 6%) or even absent (H3) in Africa. However, H3, which had a frequency of 36% and 20% in Europe and East Asia, respectively, is just a one-step mutation (SNP 871) from H2, the most frequent haplotype in the African sample. The most in-

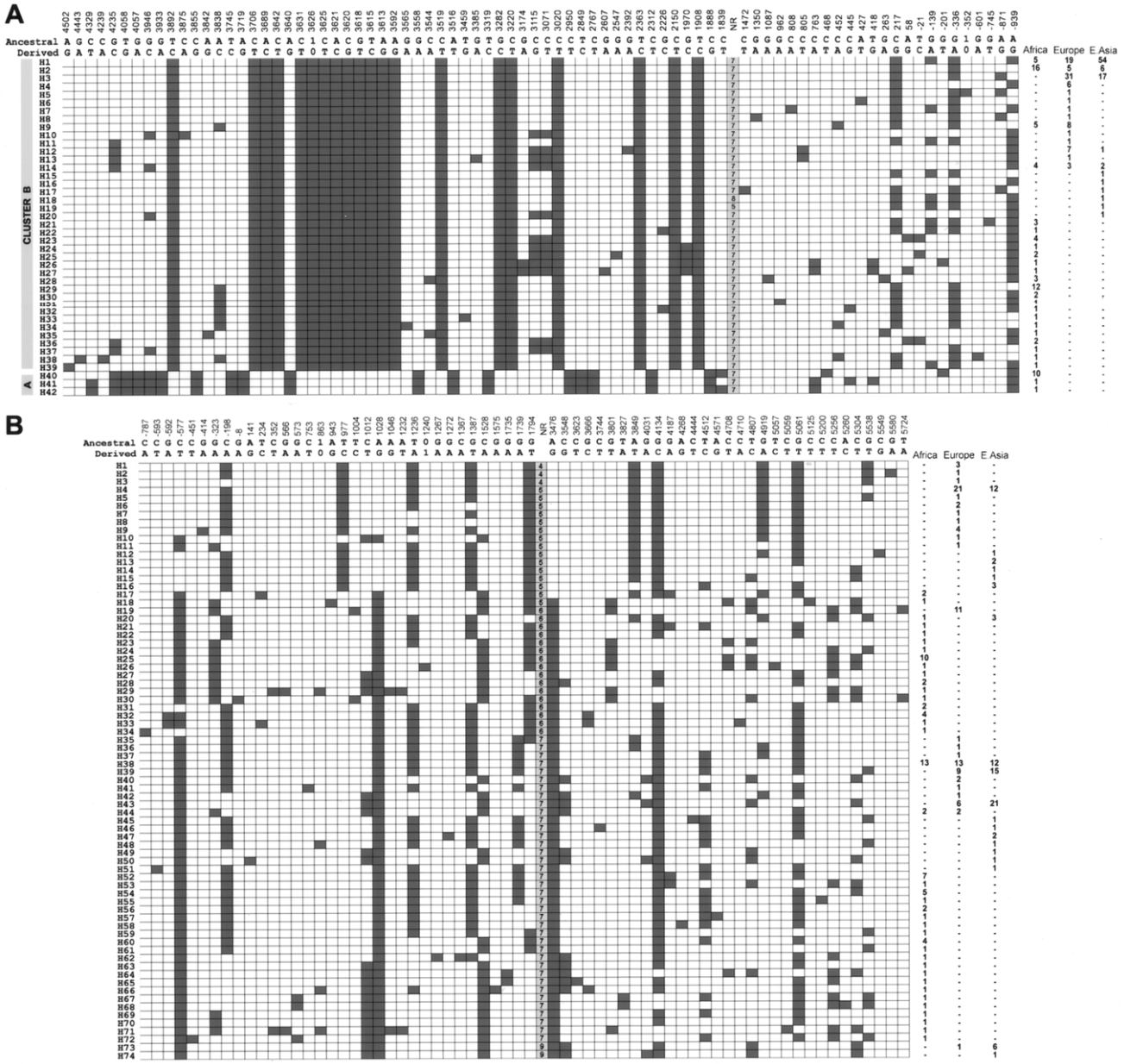

Figure 2 Inferred haplotypes for *CD209* (*A*) and *CD209L* (*B*). The chimpanzee sequence was used to deduce the ancestral state at each position, except for the *CD209L* positions 1232, 1236, and 1240. For those polymorphisms, the ancestral state was considered to be the most frequent allele. Dark boxes correspond to the derived state at each position. The numbers on the right of the figure indicate the absolute frequency of each haplotype in the different populations studied. Repeat-number variation in the neck region of each gene is reported in the gray columns with the column heads "NR." Indel polymorphisms are referred as to "1" for insertion and "0" for deletion.

teresting observation of the *CD209* haplotype variability was the presence of a highly divergent haplotype cluster. This cluster, which contains haplotypes 40–42 (referred to here as "cluster A"), differs from all other haplotypes (referred to here as "cluster B") by 35 fixed positions (fig. 2*A*). Cluster A is Africa specific and is present at a frequency of ~15%, whereas cluster B is present in the remaining African and all non-African samples. It is worth noting that three (SNPs 1839, 1888, and 1908) of the five nonsynonymous mutations identified for this gene are unique to cluster A. In all cases, these three mutations were segregating together, with the exception of one haplotype, H41, which does not contain the SNP 1839. Samples from cluster A are geographically wide-

Table 2

Summary of Diversity Indexes and Sequence-Based Neutrality Tests in the Study Populations

Gene and Population	No. of Chromosomes	No. of Segregating Sites	No. of Haplotypes	HD[a] ± SD	π[b] ± SD	θ_w[c]	Tajima's *D*	Fay and Wu's *H*
CD209:								
African	82	70	26	91.8 ± 1.6	26 ± 3.8	25.3	−.05	***−19.45***[d]
European	86	18	14	79.6 ± 3.0	6.4 ± .6	6.5	−.04	−.26
East Asian	86	12	11	56.7 ± 5.5	3.3 ± .5	4.3	−.65	−3.82[d]
Total	254	79	42	84.5 ± 1.6	13 ± 1.7	23.3		
CD209L:								
African	82	51	40	94.9 ± 1.2	16.1 ± .9	18.7	−.49	−1.52
European	86	29	23	88.8 ± 1.9	17.7 ± 1.0	10.5	***2.01***[e]	−.61
East Asian	86	27	19	86.4 ± 1.8	16.0 ± .5	9.8	***1.85***[d]	−.43
Total	254	63	74	93.6 ± .7	17.7 ± .5	18.8		

NOTE.—The values shown in bold italics correspond to significant values for both the coalescence simulation and the empirical distribution (see the "Material and Methods" section). The analyses considered a total of 5,500 and 5,391 nucleotides for *CD209* and *CD209L,* respectively.

[a] HD = haplotype diversity (%).

[b] Nucleotide diversity per base pair ($\times 10^{-4}$).

[c] Watterson's estimator per base pair ($\times 10^{-4}$).

[d] $.02 < P \leq .05$.

[e] $P \leq .02$.

spread over the entire African continent (i.e., two San from Namibia, three Bantus from Gabon and two from South Africa, three Yorubans from Nigeria, and two Mandenka from Senegal). For *CD209L,* 74 different haplotypes were observed (fig. 2*B*), with an overall heterozygosity of 94% (table 2). Only one haplotype (H38) at a frequency of ~15% was shared in the three continental regions.

To assess the degree of population differentiation, if any, we computed Wright's F_{ST} (Wright 1931), using haplotype frequencies. F_{ST} estimates were significant ($P < .0001$) for all population comparisons, indicating continental differentiation for both *CD209* and *CD209L*. However, substantial differences were observed between the two genes: the overall F_{ST} for *CD209* among Africans, Europeans, and East-Asians was 0.15, whereas *CD209L* presented a threefold lower F_{ST} value of 0.05. For both genes, the larger F_{ST} values were observed between African and East Asian populations, with F_{ST} values of 0.22 for *CD209* and 0.07 for *CD209L*.

Levels of Polymorphism and Divergence between Humans and Chimpanzees

The average nucleotide diversity (π) was strikingly different, both between the two genes and among populations (table 2). Globally, π values were three- to fivefold lower for *CD209* ($3–7 \times 10^{-4}$) than for *CD209L* ($\sim 16 \times 10^{-4}$), except for African populations, for whom the *CD209* π value was unusually high (26×10^{-4}) because of the presence of the highly divergent cluster A. Indeed, when cluster A was excluded from the analysis, the African π value dropped to 8×10^{-4}. To estimate the substitution rate of each region and evince possible mutational differences that could explain the strong contrast observed in nucleotide-diversity patterns, we determined the human-chimpanzee divergence for both genes. The average net number of differences between the two species was 77.3 substitutions (or 0.0157 substitutions per nucleotide) for *CD209* and 90.6 substitutions (or 0.0171 substitutions per nucleotide) for *CD209L*. Since the human-chimpanzee speciation occurred 5 MYA, we obtained similar nucleotide-substitution rates per site per year (*CD209,* 1.57×10^{-9}; *CD209L,* 1.70×10^{-9}).

LD

To assess the patterns of LD in the *CD209/CD209L* region, haplotypes for the entire genomic region were reconstructed using markers with an MAF of 10%. D' measures among these markers were estimated for African and non-African populations independently; the graphical representation of LD levels is illustrated in figure 3. Two distinct regions, which correspond to either *CD209* or *CD209L,* showed strong LD and are separated by a boundary that corresponds to the intergenic region. For *CD209,* a block of intragenic LD was observed in both African and non-African populations. For the African sample, 89% of all pairwise comparisons indicated significant levels of LD, whereas, for non-Africans, all D' pairwise comparisons were significant. The magnitude of intragenic recombination (and/or gene conversion) of *CD209L* was slightly higher than for *CD209*. Nevertheless, considerable and significant levels of LD were observed between sites: 83% of all LD pairwise comparisons were significant in the African group, and 99% were in the non-African sample. Overall, *CD209* exhibited a blocklike structure in both groups, whereas *CD209L* presented lower—although mostly

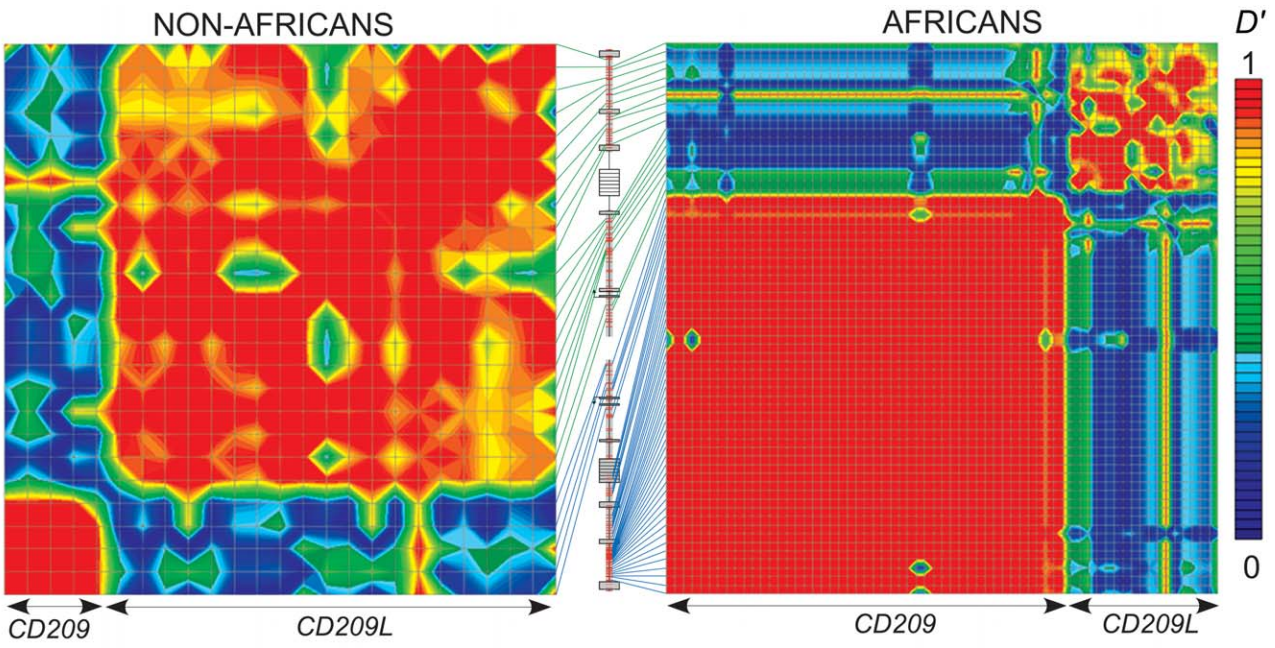

Figure 3 Pairwise *D'* LD plots in non-African and African populations. European and East Asian samples were plotted together as "non-Africans" because they showed similar levels of LD (data not shown). Red tags indicate the physical position of each SNP across the genomic region studied. Blue and green lines label the SNPs (MAF > 10%) used for *CD209* and *CD209L*, respectively, in the LD plot. For *CD209*, 47 SNPs presented an MAF > 10% in the African sample and 5 in the non-African, whereas, for *CD209L*, 18 SNPs showed an MAF > 10% in Africans and 20 in non-Africans. The high prevalence of SNPs with MAF > 10% for *CD209* in Africa is due to the presence of the highly divergent cluster A, which presents 35 diagnostic variants with a frequency of 15%.

significant—LD levels, in particular among the non-African sample.

The strong decay in LD observed in the intergenic region (fig. 3), which spans only ~14 kb, suggests the occurrence of a number of recombination events. To test the hypothesis of a possible recombination hotspot situated within this region, recombination parameters across the entire *CD209/CD209L* region (~26 kb) were computed for the three populations, by use of the recombination model implemented in Phase (v.2.1.1) (fig. 4). This model (Stephens and Donnelly 2003) estimates the position and relative intensity of the hotspot (λ) as compared with the background population recombination rate (ρ) (see the "Material and Methods" section). A λ value of 1 corresponds to absence of recombination-rate variation, whereas λ values >1 indicate the presence of a hotspot. The model detected the occurrence of a hotspot in the intergenic region, with Africans presenting a λ of 18, whereas Europeans and East Asians exhibited λ values of 63 and 53, respectively (fig. 4). We estimated the posterior probabilities of a hotspot of any kind, $\Pr(\lambda > 1)$, and of at least 10 times the background recombination rate, $\Pr(\lambda > 10)$. $\Pr(\lambda > 1)$ was 100% for all population groups, and $\Pr(\lambda > 10)$ was 64% for Africans, 97% for Europeans, and 92% for East Asians. Thus, our data clearly indicate a relative increase of the recombination levels between the two genes, which suggests the occurrence of a hotspot of recombination, the magnitude of which varies among the major ethnic groups. However, our data do not include intergenic SNPs; therefore, the exact location and width of the recombination hotspot within the intergenic region remains unclear, since this observation would be consistent with either an intense narrow hotspot or a weaker but wider hotspot.

Neutrality Tests

The identification of a strong decay in LD between *CD209* and *CD209L* facilitated the interpretation of neutrality tests, because the noise introduced by hitchhiking effects between the genes is reduced. We applied Tajima's *D* and Fay and Wu's *H* tests to determine whether these statistics significantly deviated from expectations under neutrality, using both coalescent simulations and the empirical distribution obtained from Akey et al. (2004). Globally, Tajima's *D* test indicated different tendencies for the two genes (table 2). *CD209* always yielded negative values for Tajima's *D* but never achieved significance to reject the hypothesis of neutrality, whereas *CD209L* yielded significantly positive values for non-African populations, with use of both

coalescent simulations and the empirical distribution. For Fay and Wu's *H* test, the hypothesis of neutrality was rejected for *CD209* in the African and East Asian samples (table 2).

To evaluate the selective pressures at the protein level, we performed two interspecies tests: K_A/K_S, which gives the ratio of nonsynonymous and synonymous changes between species, and the McDonald-Kreitman test, which tests the null hypothesis that the ratio of the number of fixed differences to polymorphisms is the same for both nonsynonymous and synonymous mutations. For the K_A/K_S test, *CD209* and *CD209L* showed similar values, 0.34 and 0.37, respectively. For the McDonald-Kreitman test, the hypothesis of neutrality was rejected for only *CD209*, because of a clear lack of nonsynonymous polymorphic sites (table 3).

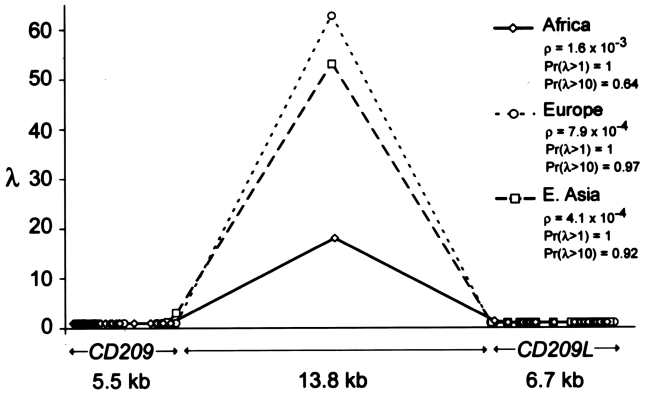

Figure 4 Estimates of the hotspot intensity (λ) for Africans, Europeans, and East Asians. Estimates of the population recombination rate (ρ) for each population as well as the posterior probabilities of $\lambda > 1$ and $\lambda > 10$ are also reported in the key.

Neck-Region Length Variation in Worldwide Populations

The identical genomic organization of *CD209* and *CD209L* is extended to the neck region, which, in both genes, encodes a track of seven coding repeats of 23 aa each (fig. 1) (Soilleux et al. 2000). A previous study has shown that the length of the neck region of *CD209L* varied between individuals of European descent (Bashirova et al. 2001). To investigate the degree of polymorphism of the neck region in both *CD209* and *CD209L*, we genotyped it in the entire HGDP-CEPH panel (1,064 individuals from 52 worldwide populations). Striking differences were observed between the two genes (see fig. 5 and table 4 for detailed allele frequencies in each population). For *CD209*, virtually no variation was observed, and the 7-repeat allele accounted for 99% of the total variability. Despite this limited variation, eight different alleles were observed, with an allele size range of 2–10 repeats, not including a 9-repeat allele. The geographic region that presented the highest variability was the Middle East, with five of the eight different alleles observed (fig. 5*A* and table 4). For *CD209L*, a completely different pattern emerged, with strong variation in allelic frequencies of different repeat numbers. Of the seven alleles observed (from 4–10-repeat allele size classes), the three most common overall were the 7- (57.42%), the 5- (23.92%), and the 6- (11.37%) repeat alleles. European, Asian, and Pacific populations presented a mosaic composition of different allelic classes, whereas 7- and 6-repeat alleles accounted for most (96%) of the African diversity (fig. 5*B*). The strong difference in the neck-region lengths between the two genes was consequently visible in the heterozygosity values: *CD209* exhibited an overall heterozygosity of only 2%, whereas *CD209L* presented a value of 54% (table 5). Our results showed that the levels of heterozygosity observed at *CD209* were considerably lower than expected, regardless of the mutation model considered (i.e., Infinite Site or Stepwise Mutation Models) (table 5). In strong contrast, although not statistically significant for individual populations, *CD209L* exhibited a pattern of an excess of heterozygosity in all populations.

Table 3

McDonald-Kreitman Test Results

Gene and Type of Site	No. of Substitutions and *P* Value for					
	Exonic Region Only			Entire Sequence[a]		
	Synonymous	Nonsynonymous	*P*	Synonymous	Nonsynonymous	*P*
CD209:			***.04***			***.009***
Fixed	4	5		51	5	
Polymorphic	6	0		86	0	
CD209L:			.23			1
Fixed	5	6		78	6	
Polymorphic	0	4		65	4	

Note.—The highly variable exon 4 has been excluded from this analysis, because no ancestral state could be inferred. Significant *P* values are shown in bold italics.

[a] Mutations in introns are considered synonymous.

Time of the Most Recent Common Ancestor for CD209

The low levels of intragenic recombination observed in *CD209* allowed maximum-likelihood coalescent analysis (Griffiths and Tavare 1994) for estimation of the time scale of the origin and evolution of this gene. Since this method assumes an infinite-site model without recombination, the same analysis for *CD209L* was not conducted because of the substantial amount of recombinant haplotypes observed. For *CD209,* only 29 of the 254 chromosomes analyzed had to be excluded, as did a single segregating site (SNP 939). The resulting *CD209* gene tree estimate, rooted with the chimpanzee sequence (i.e., the chimpanzee sequence was used to define ancestral/derived status of human mutations), is shown in figure 6. The tree is partitioned into two deep branches that correspond to haplotype clusters A and B. African samples were observed in both sides of the deepest node of the tree (i.e., in both clusters A and B), whereas non-African samples are restricted to one branch of the tree (i.e., cluster B). The maximum-likelihood estimate of θ (θ_{ML}) for *CD209* was 8.4. On the basis of this θ_{ML} value and the estimated mutation rate (1.54×10^{-4} per gene per generation), the effective population size (N_e) was 13,636, a value comparable to most figures reported in the literature (for a review, see Tishkoff and Verrelli [2003]). The T_{MRCA} of the *CD209* tree was then estimated at 2.8 ± 0.22 MYA, one of the oldest T_{MRCA} values estimated so far in the human genome (Excoffier 2002).

Table 4

Allele Relative Frequencies of Neck-Region Repeat Variation in *CD209* and *CD209L* in Individual Populations

The table is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Discussion

The *CD209/CD209L* region possesses a number of characteristics that make it a powerful tool for evolutionary inference. These two genes are not in LD, despite their very close physical vicinity (~15 kb), and each of them behaves as an independent genetic entity. Moreover, our results suggest that the *CD209/CD209L* region is a uniform landscape of genomic forces, since the two lectin-coding genes present similar mutation rates, as well as high nucleotide identity and conserved exon-intron organization (fig. 1).

Contrasting Patterns of Diversity in the CD209/CD209L *Region*

Our diversity study revealed completely different patterns for the two genes. First, levels of nucleotide diversity (π) were found to be much lower for *CD209* than for *CD209L* (table 2). On the basis of 1.42 million SNPs, the International SNP Map Working Group defined 7.5×10^{-4} as the average value of nucleotide diversity for the human genome and showed that 95% of all bins presented π values varying from 2.0×10^{-4} to

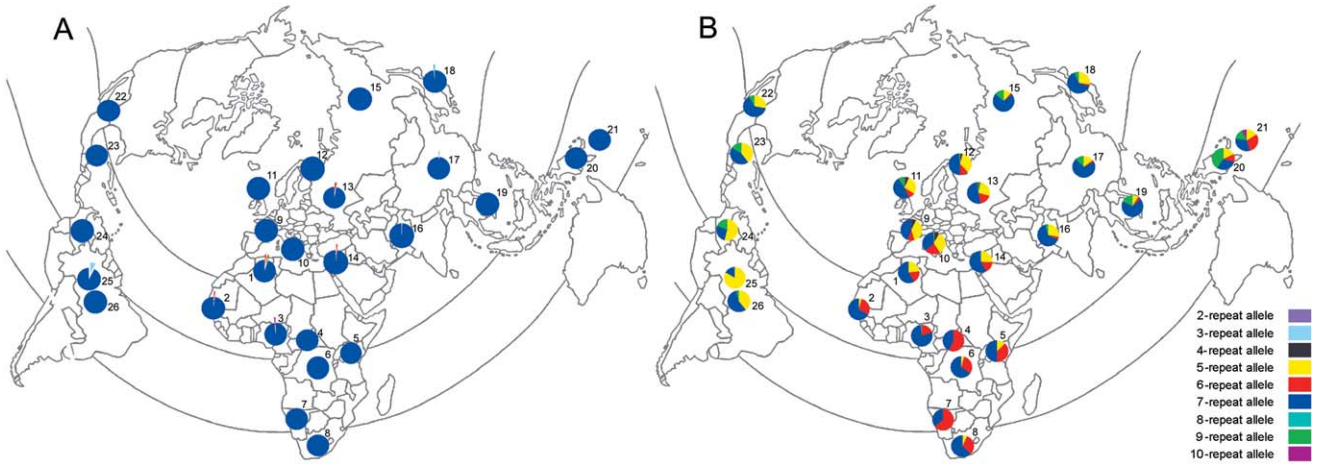

Figure 5 Geographical distribution of the neck-region repeat variation in *CD209* (*A*) and *CD209L* (*B*). Population codes are (1) Algerians; (2) Mandenka; (3) Yoruba; (4) Biaka Pygmies; (5) Northeastern Bantu from Kenya; (6) Mbuti Pygmies; (7) San; (8) South African Bantu southeastern/southwestern; (9) French and Basque from France; (10) Italian composite from Bergamo, Tuscany, and Sardinia; (11) Orcadian; (12) Russians; (13) Adygei; (14) Middle Eastern composite sample of Druze, Palestinian, and Bedouin; (15) Yakut; (16) Pakistani composite sample; (17) Chinese composite sample; (18) Japanese; (19) Cambodian; (20) Papuan; (21) Melanesian; (22) Pima; (23) Maya; (24) Piapoco and Curripaco; (25) Surui; and (26) Karitiana. For populations 16 and 17, we have pooled the different Pakistani and Chinese individual populations, respectively. For population details of these two composite groups, see the HGDP-CEPH Web site.

Table 5

Observed and Expected Heterozygosities for the Number of Repeats in the Neck Regions of *CD209* and *CD209L*

	Findings for Neck Regions of							
	CD209				*CD209L*			
	Heterozygosity		*P*		Heterozygosity		*P*	
Population	Observed	Expected[a]	ISM[b]	SMM[c]	Observed	Expected[a]	ISM[b]	SMM[c]
African	1.6	27.9	***.030***	***.000***	50	37	.328	.229
European	.6	15.3	.158	.094	62	44	.179	.304
Middle Eastern	5.6	43.1	***.018***	***.000***	61	49	.299	.095
Central/South Asian	1.4	35.1	***.003***	***.000***	52	43	.387	.098
East Asian	1.2	34.5	***.003***	***.000***	47	42	.472	.054
Oceanian	.0	…	…	…	72	53	.071	.337
American	2.8	16.3	.323	.205	45	29	.273	.440
Total sample	2.0	49.7	***.002***	***.000***	54	47	.405	***.013***

Note.—We presented only the expected heterozygosity under the infinite-site model, because no evidence for recurrent mutations were observed in our data, as suggested by the composite *CD209L* haplotypes that included the repeat variation (fig. 2), as well as by the median-joining networks (results not shown). Significant *P* values are shown in bold italics.

[a] Under the infinite-site model.

[b] Probability of the observed heterozygosity under the infinite-site model.

[c] Probability of the observed heterozygosity under the stepwise mutational model.

15.8×10^{-4} (Sachidanandam et al. 2001). In addition, an independent study analyzed nucleotide and haplotype diversity for 313 genes and defined the average π value as 5.4×10^{-4} (Stephens et al. 2001). In this context, the values observed for *CD209* ($3–7 \times 10^{-4}$) are in agreement with these genome estimations, with the exception of the African sample, which showed extreme levels of diversity (26.0×10^{-4}) because of the presence of cluster A. By contrast, the π values observed for *CD209L* ($16–18 \times 10^{-4}$) are at least twofold higher than average genome estimates and fall into the upper limit of the 95% CI defined by the SNP Consortium (Sachidanandam et al. 2001). This contrast in nucleotide diversity between the two genes can be explained either by a disparity in local mutation rates or by actual differences in selective pressures. However, no major differences in mutation rates (1.57×10^{-9} vs. 1.70×10^{-9}) were observed between the two homologues, nor was there substantial variation in GC content, which has been positively correlated with mutation rates and levels of polymorphisms (Sachidanandam et al. 2001; Smith et al. 2002; Waterston et al. 2002; Hellmann et al. 2003). Indeed, the GC content for *CD209* (53.7%) was slightly higher than that observed for *CD209L* (50.9%), which reinforces the idea that different selective pressures may indeed have been the driving force behind the distinct patterns of diversity observed. Second, the patterns of repeat variation in the neck region also turned out to be strikingly different between the two genes. *CD209* showed levels of heterozygosity of only 2%, whereas *CD209L* presented an extraordinarily high level of worldwide diversity, with an overall heterozygosity of 54% (table 5 and fig. 5). Although the neck regions of both genes share 92% of nucleotide identity, nonuniform mutation rates could, again, explain the patterns observed. However, this does not seem to be the case, since mutation-rate variation should influence the number of alleles observed rather than their frequencies, which are subject either to genetic drift or to natural selection. Indeed, we observed an even higher number of repeat alleles for *CD209* (eight alleles) than for *CD209L* (seven alleles) (table 4 and fig. 5). Overall, differences in genomic forces seem to be insufficient to explain the contrasting patterns observed at both the sequence and neck-region length variation levels; therefore, the action of differential selective pressures acting on these genes becomes the most plausible scenario.

CD209*: The Signature of a Functional Constraint*

For *CD209*, not only nucleotide diversity but also F_{ST} intercontinental values (0.15) were in conformity with previous worldwide estimations (Harpending and Rogers 2000; Akey et al. 2002; Cavalli-Sforza and Feldman 2003). For frequency-spectrum–based tests, only Fay and Wu's *H* test detected an excess of highly frequently derived alleles for the African and East Asian samples, a picture that may be interpreted as the result of a selective sweep. However, the significantly negative value observed in Africa is, again, exclusively due to the presence of cluster A, since 22 of the 35 fixed SNPs distinguishing it from cluster B corresponded to the derived allelic status in the latter cluster. Because cluster B accounts for 85% of the African variability, a clear excess

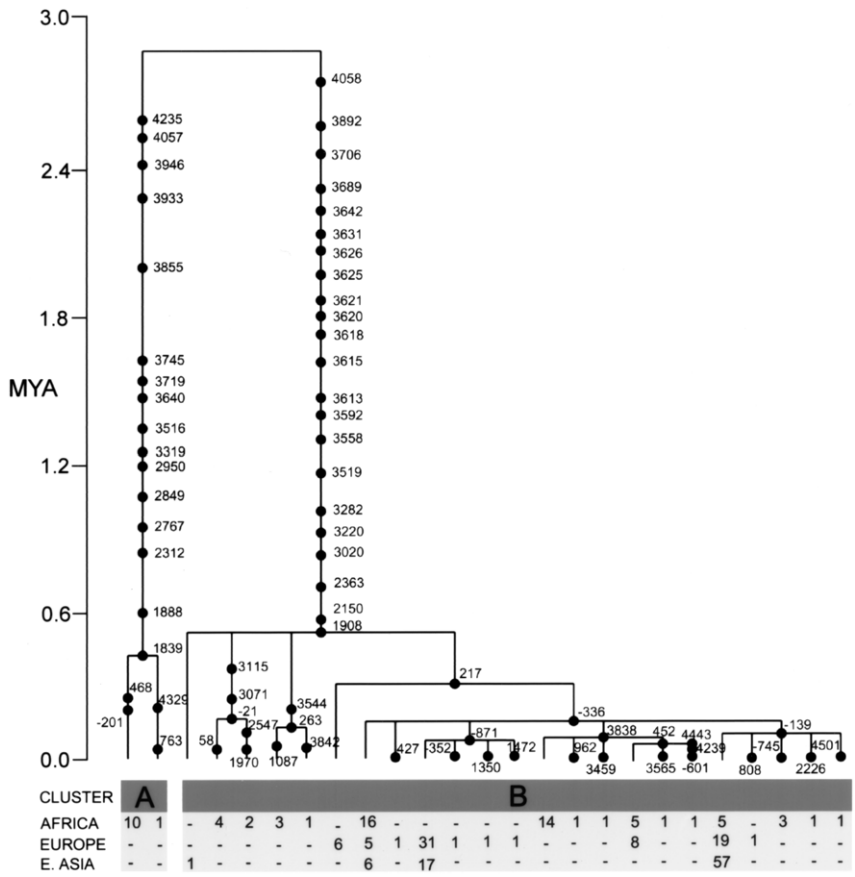

Figure 6 *CD209* estimated gene tree. Time scale is in MYA. Mutations are represented as black dots and are named for their physical position along *CD209*. For branches with multiple mutations, order in time is arbitrary. Lineage absolute frequencies in Africa, Europe, and East Asia are reported.

of frequently derived alleles was observed. The extent to which the presence of this cluster is due to either natural selection or population structure will be discussed in detail below. For East Asia, the significance of the *H* test is also questionable when accounting for the confounding effects of demography. Indeed, when we plotted our *H* value against the empirical distribution of 132 *H* values from non-African populations (Akey et al. 2004), the East Asian *P* value became nonsignificant ($P = .36$). This observation reinforces the idea that the *H* test is particularly sensitive to past bottlenecks and/or population subdivision (Przeworski 2002). Thus, regarding the global levels of sequence diversity, the *CD209* locus seems to evolve under evolutionary neutrality. Nevertheless, when we focused our analyses at the protein level, signs of natural selection were uncovered. Indeed, the McDonald-Kreitman test rejected neutrality for this gene because of a clear excess of polymorphic synonymous sites (i.e., a lack of nonsynonymous variants). In addition, when the number of synonymous sites (146) versus nonsynonymous sites (499) was compared with the observed number of synonymous (5) versus nonsynonymous (0) mutations, we detected a significant lack of nonsynonymous mutations (two-tailed Fisher exact test, $P = 6.3 \times 10^{-4}$). These observations point to a strong selective constraint acting on *CD209* that prevents the accumulation of amino acid replacements over time.

Further support for a functional constraint in *CD209* comes from the patterns of diversity observed in the neck region. In contrast to *CD209L,* virtually no variation was observed at *CD209* (fig. 5*A*), with the 7-repeat allele accounting for 99% of the total variability. Moreover, the low levels of heterozygosity observed resulted in a consistent rejection of mutation-drift equilibrium in almost all geographical regions (table 5). The probability of finding such a low heterozygosity value, given the overall number of alleles observed, was estimated to be <0.2%, independent of the mutational model considered (table 5). Thus, the fact that no alleles other than the 7-repeat allele have increased in frequency, together with recent studies addressing the functional consequences of

Table 6

AMOVA for the Neck Region of *CD209L*

Sample[a]	No. of Regions	No. of Populations	AMOVA Value (95% CI) Inferred for *CD209L*[b]		
			Within Populations	Among Populations within Regions	Among Regions
World	7	52	90.4 (93.8–94.3)	2.1 (2.3–2.5)	7.57 (3.3–3.9)
Africa	1	6	93.9 (96.7–97.1)	6.1 (2.9–3.3)	
Eurasia	3	21	97.0 (98.2–98.4)	.2 (1.1–1.3)	2.8 (.4–.6)
Europe	1	8	99.5 (99.1–99.4)	.5 (0.6–0.9)	
Middle-East	1	4	100 (98.6–98.8)	0 (1.2–1.4)	
Central/South Asia	1	9	99.5 (98.5–98.8)	.5 (1.2–1.5)	
East Asia	1	18	99.3 (98.6–98.9)	.7 (1.1–1.4)	
Oceania	1	2	96.0 (92.8–94.3)	4.0 (5.7–7.2)	
America	1	5	86.7 (87.7–89)	13.3 (11.0–12.3)	

Note.—No comparisons were performed for the *CD209* neck region, because virtually no variation was observed at that locus.

[a] Populations are grouped as described by Rosenberg et al. (2002).

[b] AMOVA values are from our *CD209L* study; 95% CIs are defined from 377 autosomal microsatellites in the same population panel (Rosenberg et al. 2002).

repeat-number variation in this region (Bernhard et al. 2004; Feinberg et al. 2005), strongly suggests a clear reduced fitness of any allele other than the 7-repeat allele. Interestingly, it has been recently shown that a protein with two fewer repeats (a 5-repeat allele) results in a partial dissociation of the final tetramer, whereas a protein with <5 repeats exhibits a dramatic reduction in overall stability (Feinberg et al. 2005), with all these differences having a direct impact on the quality of ligand-binding functions (Bernhard et al. 2004). Taken together, the patterns of diversity observed at *CD209* clearly point to a strong functional constraint acting on this gene and further support the proposed crucial role of this lectin in pathogen recognition and in the early steps of immune response (Geijtenbeek et al. 2000*b*, 2004).

CD209L: *Relaxation of the Functional Constraint or Balancing Selection?*

In clear contrast to its homologue, *CD209L* presented extremely elevated nucleotide-diversity levels. High levels of diversity can result either from a relaxation of the functional constraint, which allows the stochastic accumulation of new mutations, or from the action of balancing selection, which maintains over time two or more functionally different alleles (and all linked variation) at intermediate frequencies. Several lines of evidence lend support to the selective hypothesis. First, if *CD209L* nucleotide diversity has been driven by the action of balancing selection, population-genetics relationships would have been accordingly altered. In this context, diversity studies in neutral, or assumedly neutral, regions of the genome—such as the Y chromosome (Underhill et al. 2000; Hammer et al. 2001; Jobling and Tyler-Smith 2003), mtDNA (Wallace et al. 1999; Ingman et al. 2000; Mishmar et al. 2003), *Alu* insertions (Watkins et al. 2001), as well as some autosomal genes (Stephens et al. 2001; Akey et al. 2004)—showed that African populations are genetically more diverse than are non-Africans, an observation generally interpreted as a support of the "Out of Africa" model for the origin of modern humans (Lewin 1987). For *CD209L,* even if we observed 1.5 times more segregating sites in African than in non-African populations, as indicated by the higher θ_w value found in Africa, similar values of nucleotide diversity were detected in the three groups, with Europeans presenting even higher π values than do Africans. This unusual scenario, which is at odds with neutral expectations, has already been described for other regions of the genome, such as the *β-globin* gene and the *5′ cis*-regulatory region of *CCR5,* for which the action of balancing selection has been convincingly proposed (Harding et al. 1997; Bamshad et al. 2002). Second, balancing selection tends to increase within-population diversity while decreasing F_{ST}, compared with neutrally evolving loci (Cavalli-Sforza 1966; Harpending and Rogers 2000; Akey et al. 2002; Bamshad and Wooding 2003; Cavalli-Sforza and Feldman 2003). Indeed, our data are compatible with these predictions, since the 5% F_{ST} value observed for *CD209L* is threefold lower than that estimated for *CD209* (15%) and is similar to that found, for example, for the bitter-taste receptor gene (5.6%), for which there is compelling evidence of balancing-selection action (Wooding et al. 2004). Third, results of our Tajima's *D* analysis were significantly positive for European and East Asian populations, because of the skew of *CD209L* frequency spectrum toward an excess of intermediate-frequency alleles (table 2), a pattern that further supports the action of balancing selection. However, since the null model used to assess

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 7 Coalescent-based simulations (2×10^4) of the expected T_{MRCA} distribution of *CD209*.

significance makes unrealistic assumptions about past population demography (i.e., constant population sizes), the rejection of the standard neutral model cannot be interpreted as unambiguous evidence of selection. Indeed, the observation that only non-African populations showed a significant departure from neutrality raises the question of whether these patterns could have resulted instead from the bottleneck that occurred during the Out of Africa exodus. A way to circumvent this conundrum is to analytically integrate the fact that demography affects all the genome equally, whereas selection directs its effects toward specific loci. Thus, to correct for the confounding effects of demography, we plotted our results against the empirical distributions of Akey et al. (2004) for Tajima's *D* statistics. Our values remained significant for *CD209L,* which therefore reinforces the idea that the pattern observed is unlikely to be the sole result of demography.

Last, if the patterns of variation in *CD209L* represent the molecular signature of balancing selection, at least in non-Africans, then a functional target of such selective regime is needed. In this context, the neck region constitutes an excellent candidate, since it plays a major mediating role in the orientation and flexibility of the carbohydrate-recognition domain. Since this domain is directly involved in pathogen recognition, neck-region length variation has important consequences for the pathogen-binding properties of these lectins (Mitchell et al. 2001; Bernhard et al. 2004; Feinberg et al. 2005). In perfect agreement with the results of our sequence-based data set, higher diversity in repeat variation was observed in the neck region among non-African populations (Native Americans excepted). Out of Africa, at least three alleles account for most population diversity, whereas, in Africa, the 6- and 7-repeat alleles alone account for 96% of the global variability (fig. 5*B*). Again, the higher diversity observed out of Africa could be due to a higher level of relaxation of the functional constraint of the neck region in non-African compared with African populations, which would lead to a random accumulation of proteins with varying neck-region lengths among non-Africans. Conversely, these patterns could also be explained by the action of balancing selection in non-Africans and could therefore point to the neck region as the functional target of such selective regime. To evaluate the plausibility of these two conflicting scenarios, we compared the variation in the *CD209L* neck region with that inferred from 377 neutral autosomal microsatellites typed elsewhere for the same population panel (Rosenberg et al. 2002). We reasoned that if *CD209L* diversity has been shaped only by demography (i.e., bottleneck out of Africa), the distribution of genetic variance at different hierarchical levels should be comparable to that inferred through the neutral markers. On the other hand, if selection has driven the *CD209L* neck-region diversity, population-genetics distances would be influenced accordingly and would therefore differ from neutral expectations. Indeed, the AMOVA values inferred for *CD209L* fell systematically outside the 95% CI defined for the microsatellite data set (table 6). We observed that populations within Europe, Asia, the Middle East, and Oceania exhibited lower-than-expected diversity among populations within the same region. A reduction of genetic distances between populations is expected under balancing selection; therefore, the results from the *CD209L* neck region favor, once again, the action of this selective regime in most non-African populations, in detriment of the neutral hypothesis. One may argue that the differences in the proportions of genetic variance between our data and those of Rosenberg et al. (2002) could be due to differences in the pace of mutation between microsatellite loci and our neck repeated region that could be considered a "coding minisatellite." However, under neutrality, differences in mutation rate should have a similar and proportional effect in all population comparisons and should influence all values with a similar tendency (i.e., higher or lower values). Indeed, this is not the case: populations within Europe, the Middle East, Central/South Asia, East Asia, and Oceania turned out to be genetically closer than expected, whereas populations within Africa and the Americas exhibited the opposite pattern (table 6), which makes it highly unlikely that mutation-rate differences influenced our conclusions.

Taken together, the integration of the results from levels of nucleotide and amino acid diversity, neutrality tests, population-genetics distances, and neck-region length variation in *CD209* and *CD209L* clearly points to a situation in which *CD209* has been under a strong selective constraint that prevents accumulation of any of amino acid changes over time, whereas *CD209L* variability has most likely been driven by the action of balancing selection, at least in non-African populations.

The Footprints of Ancestral Population Diversity

In apparent dichotomy with the strong selective constraint described for *CD209,* we observed an unusual excess of diversity of 35 fixed differences separating the two basal branches of the gene tree (fig. 6). In addition, we estimated a T_{MRCA} of 2.8 ± 0.22 MYA, a time that places the most recent common ancestor of *CD209* back

in the Pliocene epoch, before the estimated time for the origins of the genus *Homo* ~1.9 MYA (Wood 1996; Wood and Collard 1999). A number of studies have already reported loci that present unusually deep coalescent times (Harris and Hey 1999; Zhao et al. 2000; Webster et al. 2003; Garrigan et al. 2005*a,* 2005*b*), but our estimation for *CD209* remains one of the deepest T_{MRCA} values yet reported (Excoffier 2002). The probability of finding such a deep coalescence time under a scenario of a random-mating population was estimated, through a coalescent process (Laval and Excoffier 2004), to be very low ($P = .018$) (see fig. 7). In addition to the unexpected antiquity of the *CD209* locus, we observed a peculiar tree topology made of two highly divergent and frequency-unbalanced lineages, cluster A embracing only 2 internal haplotypes and cluster B comprising the remaining 23 (fig. 6).

Different hypotheses can account for such elongated and divergent haplotype patterns. Indeed, the high levels of nucleotide identity between *CD209* and *CD209L* could have led to gene conversion between the two genes, an event that would explain the outlier position of cluster A in the context of *CD209* phylogeny. We reasoned that if gene conversion has occurred, we expect that the derived alleles distinguishing clusters A and B in *CD209* would correspond to the allelic state observed in their homologous positions in *CD209L*. Of all positions, only four fit this criterion. In addition, these positions were not physically clustered, which therefore excludes a major gene-conversion event as the explanation of the divergent *CD209* phylogeny.

Two other circumstances may be responsible for the topology and the time depth of the *CD209* gene tree: long-standing balancing selection or ancient population structure, with Africa, in both cases, being the arena of such events (i.e., cluster A is restricted to Africa). Several lines of evidence argue against the balancing-selection hypothesis. First, under this selective regime, one would expect that Tajima's *D* test would also point in this direction by yielding significantly positive values, which is not the case (table 2). Second, such a long-standing balancing selection in Africa would have entailed a number of recombinant haplotypes between clusters A and B, which, again, is not the case, as illustrated by the high LD levels at *CD209* (fig. 3). Third, a claim of balancing selection at this locus must imply a functional difference between the two balanced alleles. Indeed, three nonsynonymous mutations, situated in the neck region, separate cluster A and B, and they could correspond to the alleles under selection. But, if the neck region is the target of selection, it is more likely that the balanced alleles would correspond to different numbers of repeats rather than punctual nucleotide variation within each track, as observed for *CD209L* and suggested by functional studies (Bernhard et al. 2004; Feinberg et al. 2005). Since no variation in the number of repeats was detected between both clusters, we predict that there are no major functional differences between the two lineages. Taken together, maintenance of ancient lineages by balancing selection does not seem to be responsible for the observed haplotype divergence. In this view, the patterns observed are best explained by an ancestral population structure on the African continent. Indeed, several studies have already proposed that African populations must have been more strongly subdivided and isolated than non-African ones (Harris and Hey 1999; Labuda et al. 2000; Excoffier 2002; Goldstein and Chikhi 2002; Harding and McVean 2004; Satta and Takahata 2004; Garrigan et al. 2005*a*). In particular, a recent study of the Xp21.1 locus presented convincing statistical evidence that supports the hypothesis that our species does not descend from a single, historically panmictic population (Garrigan et al. 2005*a*). The divergent haplotype pattern observed at the Xp21.1 locus prompted those authors to explain their data under the isolation-and-admixture (IAA) model and/or a metapopulation model (Harding and McVean 2004; Wakeley 2004). Indeed, as observed for *CD209,* under an IAA model, the two basal branches are expected to be longer than those under a Wright-Fisher model, depending on the length of time subpopulations spent in isolation. The extent to which the IAA model fits the data depends on the number of mutations, referred as to "congruent sites," occurring in the two basal branches of the genealogy. For Xp21.1, 10 congruent sites over 24 polymorphisms were observed (i.e., ~42% of the total number of sites). We applied the same approach to *CD209* and obtained a very similar percentage of ~45%, in good accordance with the IAA model. Our observations, together with a number of autosomal diversity studies, show that modern human diversity appears to have kept genetic traces of admixture among archaic hominid populations. However, a number of questions remain unanswered, such as the time when these admixture events occurred (i.e., before or after the appearance of anatomically modern humans), the precise quantitative contribution of ancient genetic material to our modern gene pool, and the geographic provenance of these genetic vestiges.

Conclusions

The need of continuous evolution for both the human host and the pathogens is predicted by the Red Queen hypothesis (Van Valen 1973; Bell 1982), in reference to the remark of the Red Queen to Alice in *Through the Looking Glass* (Carroll 1872): "Now, here, you see, it takes all the running you can do, to keep in the same place." This metaphor provides a conceptual framework for understanding how interactions between the two species lead to constant natural selection for adaptation and

counteradaptation. In this context, one feature exploited by the host immunity genes to increase their defense potential is gene duplication by retention, through conservation of one duplicate, of the currently useful function of the encoded protein, while its twin is liberated to mutate and possibly acquire novel functions (Ohno 1970; Trowsdale and Parham 2004). The lectins CD209 and CD209L represent a prototypic model of a duplicated progeny of ancestral genes that interact with a vast spectrum of pathogens. Our results clearly indicate that these duplicated genes have evolved, and might still evolve, under completely different evolutionary pressures. Whereas one, *CD209,* shows signals of strong conservation, its paralogue, *CD209L,* exhibits an excess of sequence diversity compatible with the action of balancing selection. In addition, the strong contrast observed in length variation of the neck region between the two genes may have important consequences in medical genetics. In this context, association studies are now needed that correlate length variation of the neck region and susceptibility to infectious diseases whose etiological agents are known to interact with one (or both) of these lectins.

More generally, our study has revealed that even a short segment of the human genome can help uncover an extraordinarily complex evolutionary history, including different pathogen pressures on host immunity genes, as well as traces of ancient population structure in the African continent. The coming years will certainly bring unprecedented large data sets of sequence diversity, genomewide and populationwide, with each genomic region possibly revealing a different aspect of human history. The integration of all these apparently independent pieces of the same reality will provide us with a much broader and more realistic view of the demographic history of the human species, as well as of human adaptation to the different environmental conditions imposed not only by pathogens but also by other major factors such as climate and nutritional resources.

Acknowledgments

We warmly acknowledge Guillaume Laval for useful suggestions on the use of SIMCOAL software, Laurent Excoffier and Francesca Luca for stimulating discussions, and two reviewers for constructive comments on the first version of the manuscript. L.B.B. was supported by Fundação para a Ciência e a Tecnologia fellowship SFRH/BD/18580/2004.

Web Resources

The URLs for data presented herein are as follows:

Arlequin, http://lgb.unige.ch/arlequin/

BOTTLENECK, http://www.montpellier.inra.fr/CBGP/softwares/bottleneck/bottleneck.html

Center for Statistical Genetics, http://www.sph.umich.edu/csg/abecasis/GOLD/ (for GOLD software)

Centre National de Genotypage, http://software.cng.fr/ (for GENALYS software)

DnaSP, http://www.ub.es/dnasp/

GENETREE Software, http://www.stats.ox.ac.uk/~griff/software.html

HGDP-CEPH Human Genome Diversity Cell Line Panel, http://www.cephb.fr/HGDP-CEPH-Panel/

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for dendritic cell–specific ICAM-3 grabbing nonintegrin and liver/lymph node–specific ICAM-3 grabbing nonintegrin)

Phase, http://www.stat.washington.edu/stephens/phase.html

SIMCOAL2, http://cmpg.unibe.ch/software/simcoal2/

References

Abecasis GR, Cookson WO (2000) GOLD—graphical overview of linkage disequilibrium. Bioinformatics 16:182–183

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Alvarez CP, Lasala F, Carrillo J, Muniz O, Corbi AL, Delgado R (2002) C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in *cis* and in *trans*. J Virol 76: 6841–6844

Appelmelk BJ, van Die I, van Vliet SJ, Vandenbroucke-Grauls CM, Geijtenbeek TB, van Kooyk Y (2003) Cutting edge: carbohydrate profiling identifies new pathogens that interact with dendritic cell-specific ICAM-3-grabbing nonintegrin on dendritic cells. J Immunol 170:1635–1639

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. Proc Natl Acad Sci USA 99:10539–10544

Bashirova AA, Geijtenbeek TB, van Duijnhoven GC, van Vliet SJ, Eilering JB, Martin MP, Wu L, Martin TD, Viebig N, Knolle PA, KewalRamani VN, van Kooyk Y, Carrington M (2001) A dendritic cell-specific intercellular adhesion molecule 3-grabbing nonintegrin (DC-SIGN)-related protein is highly expressed on human liver sinusoidal endothelial cells and promotes HIV-1 infection. J Exp Med 193:671–678

Bashirova AA, Wu L, Cheng J, Martin TD, Martin MP, Benveniste RE, Lifson JD, KewalRamani VN, Hughes A, Carrington M (2003) Novel member of the *CD209* (*DC-SIGN*) gene family in primates. J Virol 77:217–227

Bell G (1982) The masterpiece of nature: the evolution and genetics of sexuality. University of California Press, Berkeley

Bergman MP, Engering A, Smits HH, van Vliet SJ, van Bodegraven AA, Wirth HP, Kapsenberg ML, Vandenbroucke-

Grauls CM, van Kooyk Y, Appelmelk BJ (2004) *Helicobacter pylori* modulates the T helper cell 1/T helper cell 2 balance through phase-variable interaction between lipopolysaccharide and DC-SIGN. J Exp Med 200:979–990

Bernhard OK, Lai J, Wilkinson J, Sheil MM, Cunningham AL (2004) Proteomic analysis of DC-SIGN on dendritic cells detects tetramers required for ligand binding but no association with CD4. J Biol Chem 279:51828–51835

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al (2002) A human genome diversity cell line panel. Science 296:261–262

Carroll L (1872) Through the looking glass. Macmillan, London

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond B Biol Sci 164:362–379

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet Suppl 33:266–275

Colmenares M, Puig-Kroger A, Pello OM, Corbi AL, Rivas L (2002) Dendritic cell (DC)-specific intercellular adhesion molecule 3 (ICAM-3)-grabbing nonintegrin (DC-SIGN, CD209), a C-type surface lectin in human DCs, is a receptor for *Leishmania* amastigotes. J Biol Chem 277:36766–36769

Cook DN, Hollingsworth JW Jr, Schwartz DA (2003) Toll-like receptors and the genetics of innate immunity. Curr Opin Allergy Clin Immunol 3:523–529

Cooke GS, Hill AV (2001) Genetics of susceptibility to human infectious disease. Nat Rev Genet 2:967–977

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014

Curtis BM, Scharnowske S, Watson AJ (1992) Sequence and expression of a membrane-associated C-type lectin that exhibits CD4-independent binding of human immunodeficiency virus envelope glycoprotein gp120. Proc Natl Acad Sci USA 89:8356–8360

Excoffier L (2002) Human demographic history: refining the recent African origin model. Curr Opin Genet Dev 12:675–682

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Feinberg H, Guo Y, Mitchell DA, Drickamer K, Weis WI (2005) Extended neck regions stabilize tetramers of the receptors DC-SIGN and DC-SIGNR. J Biol Chem 280:1327–1335

Flint J, Harding RM, Boyce AJ, Clegg JB (1998) The population genetics of the haemoglobinopathies. Baillieres Clin Haematol 11:1–51

Fujita T, Matsushita M, Endo Y (2004) The lectin-complement pathway—its role in innate immunity and evolution. Immunol Rev 198:185–202

Gardner JP, Durso RJ, Arrigale RR, Donovan GP, Maddon PJ, Dragic T, Olson WC (2003) L-SIGN (CD 209L) is a liver-specific capture receptor for hepatitis C virus. Proc Natl Acad Sci USA 100:4498–4503

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF (2005*a*) Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics 170:1849–1856

Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF (2005*b*) Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol 22:189–192

Geijtenbeek TB, Kwon DS, Torensma R, van Vliet SJ, van Duijnhoven GC, Middel J, Cornelissen IL, Nottet HS, KewalRamani VN, Littman DR, Figdor CG, van Kooyk Y (2000*a*) DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. Cell 100:587–597

Geijtenbeek TB, Torensma R, van Vliet SJ, van Duijnhoven GC, Adema GJ, van Kooyk Y, Figdor CG (2000*b*) Identification of DC-SIGN, a novel dendritic cell-specific ICAM-3 receptor that supports primary immune responses. Cell 100:575–585

Geijtenbeek TB, van Vliet SJ, Engering A, 't Hart BA, van Kooyk Y (2004) Self- and nonself-recognition by C-type lectins on dendritic cells. Annu Rev Immunol 22:33–54

Geijtenbeek TB, Van Vliet SJ, Koppel EA, Sanchez-Hernandez M, Vandenbroucke-Grauls CM, Appelmelk B, Van Kooyk Y (2003) Mycobacteria target DC-SIGN to suppress dendritic cell function. J Exp Med 197:7–17

Gilad Y, Rosenberg S, Przeworski M, Lancet D, Skorecki K (2002) Evidence for positive selection and population structure at the human *MAO-A* gene. Proc Natl Acad Sci USA 99:862–867

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3:129–152

Griffiths RC, Tavare S (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344:403–410

Halary F, Amara A, Lortat-Jacob H, Messerle M, Delaunay T, Houles C, Fieschi F, Arenzana-Seisdedos F, Moreau JF, Dechanet-Merville J (2002) Human cytomegalovirus binding to DC-SIGN is required for dendritic cell infection and target cell *trans*-infection. Immunity 17:653–664

Haldane JBS (1949) Disease and evolution. Ric Sci Suppl A 8:68–76

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18:1189–1203

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789

Harding RM, McVean G (2004) A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev 14:667–674

Harpending H, Rogers A (2000) Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361–385

Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96:3320–3324

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet 72:1527–1535

Hughes AL, Hughes MK, Howell CY, Nei M (1994) Natural

selection at the class II major histocompatibility complex loci of mammals. Philos Trans R Soc Lond B Biol Sci 346:359–367

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Janeway CA Jr, Medzhitov R (2002) Innate immune recognition. Annu Rev Immunol 20:197–216

Jeffers SA, Tusell SM, Gillim-Ross L, Hemmila EM, Achenbach JE, Babcock GJ, Thomas WD Jr, Thackray LB, Young MD, Mason RJ, Ambrosino DM, Wentworth DE, Demartini JC, Holmes KV (2004) CD209L (L-SIGN) is a receptor for severe acute respiratory syndrome coronavirus. Proc Natl Acad Sci USA 101:15748–15753

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4: 598–612

Kimbrell DA, Beutler B (2001) The evolution and genetics of innate immunity. Nat Rev Genet 2:256–267

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Labuda D, Zietkiewicz E, Yotova V (2000) Archaic lineages in the history of modern humans. Genetics 156:799–808

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485–2487

Lederberg J (1999) J. B. S. Haldane (1949) on infectious disease and evolution. Genetics 153:1–3

Lewin R (1987) Africa: cradle of modern humans. Science 237: 1292–1295

Lewontin RC (1964) The interaction of selection and linkage. II. Optimum models. Genetics 50:757–782

Lozach PY, Lortat-Jacob H, de Lacroix de Lavalette A, Staropoli I, Foung S, Amara A, Houles C, Fieschi F, Schwartz O, Virelizier JL, Arenzana-Seisdedos F, Altmeyer R (2003) DC-SIGN and L-SIGN are high affinity binding receptors for hepatitis C virus glycoprotein E2. J Biol Chem 278: 20358–20366

Marzi A, Gramberg T, Simmons G, Moller P, Rennekamp AJ, Krumbiegel M, Geier M, Eisemann J, Turza N, Saunier B, Steinkasserer A, Becker S, Bates P, Hofmann H, Pohlmann S (2004) DC-SIGN and DC-SIGNR interact with the glycoprotein of Marburg virus and the S protein of severe acute respiratory syndrome coronavirus. J Virol 78:12090–12095

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654

McGreal EP, Martinez-Pomares L, Gordon S (2004) Divergent roles for C-type lectins expressed by cells of the innate immune system. Mol Immunol 41:1109–1121

Medzhitov R (2001) Toll-like receptors and innate immunity. Nat Rev Immunol 1:135–145

Medzhitov R, Janeway CA Jr (1998*a*) An ancient system of host defense. Curr Opin Immunol 10:12–15

——— (1998*b*) Innate immune recognition and control of adaptive immune responses. Semin Immunol 10:351–353

——— (2000) Innate immunity. N Engl J Med 343:338–344

——— (2002) Decoding the patterns of self and nonself by the innate immune system. Science 296:298–300

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA 100:171–176

Mitchell DA, Fadden AJ, Drickamer K (2001) A novel mechanism of carbohydrate recognition by the C-type lectins DC-SIGN and DC-SIGNR: subunit organization and binding to multivalent ligands. J Biol Chem 276:28939–28945

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Berlin

Ohta T (1991) Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. Proc Natl Acad Sci USA 88:6716–6720

Ota T, Sitnikova T, Nei M (2000) Evolution of vertebrate immunoglobulin variable gene segments. Curr Top Microbiol Immunol 248:221–245

Pohlmann S, Soilleux EJ, Baribaud F, Leslie GJ, Morris LS, Trowsdale J, Lee B, Coleman N, Doms RW (2001) DC-SIGNR, a DC-SIGN homologue expressed in endothelial cells, binds to human and simian immunodeficiency viruses and activates infection in trans. Proc Natl Acad Sci USA 98: 2670–2675

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Satta Y, Takahata N (2004) The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. Mol Ecol 13:877–886

Schneider S, Roessli D, Excoffier L (2000) Genetics and biometry laboratory, University of Geneva, Geneva

Smith NG, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12: 1350–1356

Soilleux EJ (2003) DC-SIGN (dendritic cell-specific ICAM-grabbing non-integrin) and DC-SIGN-related (DC-SIGNR): friend or foe? Clin Sci (Lond) 104:437–446

Soilleux EJ, Barten R, Trowsdale J (2000) DC-SIGN; a related gene, DC-SIGNR; and CD23 form a cluster on 19p13. J Immunol 165:2937–2942

Soilleux EJ, Morris LS, Lee B, Pohlmann S, Trowsdale J, Doms RW, Coleman N (2001) Placental expression of DC-

SIGN may mediate intrauterine vertical transmission of HIV. J Pathol 195:586–592

Soilleux EJ, Morris LS, Leslie G, Chehimi J, Luo Q, Levroney E, Trowsdale J, Montaner LJ, Doms RW, Weissman D, Coleman N, Lee B (2002) Constitutive and induced expression of DC-SIGN on dendritic cell and macrophage subpopulations in situ and in vitro. J Leukoc Biol 71:445–457

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489–493

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Tailleux L, Schwartz O, Herrmann JL, Pivert E, Jackson M, Amara A, Legres L, Dreher D, Nicod LP, Gluckman JC, Lagrange PH, Gicquel B, Neyrolles O (2003) DC-SIGN is the major *Mycobacterium tuberculosis* receptor on human dendritic cells. J Exp Med 197:121–127

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahashi M, Matsuda F, Margetic N, Lathrop M (2003) Automated identification of single nucleotide polymorphisms from sequencing data. J Bioinform Comput Biol 1:253–265

Tassaneetrithep B, Burgess TH, Granelli-Piperno A, Trumpfheller C, Finke J, Sun W, Eller MA, Pattanapanyasat K, Sarasombath S, Birx DL, Steinman RM, Schlesinger S, Marovich MA (2003) DC-SIGN (CD209) mediates dengue virus infection of human dendritic cells. J Exp Med 197:823–829

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD:* recent origin of alleles that confer malarial resistance. Science 293:455–462

Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340

Trowsdale J, Parham P (2004) Mini-review: defense strategies and immunity-related genes. Eur J Immunol 34:7–17

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet 13:R245–R254

Van Liempt E, Imberty A, Bank CM, Van Vliet SJ, Van Kooyk Y, Geijtenbeek TB, Van Die I (2004) Molecular basis of the differences in binding properties of the highly related C-type lectins DC-SIGN and L-SIGN to Lewis X trisaccharide and *Schistosoma mansoni* egg antigens. J Biol Chem 279:33161–33167

Van Valen L (1973) A new evolutionary law. Evol Theory 1: 1–30

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am J Hum Genet 71:1112–1128

Wakeley J (2004) Metapopulation models for historical inference. Mol Ecol 13:865–875

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238:211–230

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB (2001) Patterns of ancestral human diversity: an analysis of *Alu*-insertion and restriction-site polymorphisms. Am J Hum Genet 68:738–752

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Webster MT, Clegg JB, Harding RM (2003) Common 5′ β-globin RFLP haplotypes harbour a surprising level of ancestral sequence mosaicism. Hum Genet 113:123–139

Wood B (1996) Human evolution. Bioessays 18:945–954

Wood B, Collard M (1999) The human genus. Science 284: 65–71

Wooding S, Kim U-k, Bamshad MJ, Larsen J, Jorde LB, Drayna D (2004) Natural selection and molecular evolution in *PTC,* a bitter-taste receptor gene. Am J Hum Genet 74:637–646

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, Ramos-Onsins S, Yu N, Li WH (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA 97:11354–11358