Excess of Rare Amino Acid Polymorphisms in the Toll-like Receptor 4 in Humans

Irina Smirnova,*[,1] Martha T. Hamblin,[†,1] Colleen McBride,* Bruce Beutler* and Anna Di Rienzo[†]

**Department of Internal Medicine and the Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390 and [†]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637*

Manuscript received December 6, 2000
Accepted for publication May 7, 2001

ABSTRACT

The Toll-like receptor 4 protein acts as the transducing subunit of the lipopolysaccharide receptor complex and assists in the detection of Gram-negative pathogens within the mammalian host. Several lines of evidence support the view that variation at the *TLR4* locus may alter host susceptibility to Gram-negative infection or the outcome of infection. Here, we surveyed *TLR4* sequence variation in the complete coding region (2.4 kb) in 348 individuals from several population samples; in addition, a subset of the individuals was surveyed at 1.1 kb of intronic sequence. More than 90% of the chromosomes examined encoded the same structural isoform of TLR4, while the rest harbored 12 rare amino acid variants. Conversely, the variants at silent sites (intronic and synonymous positions) occur at both low and high frequencies and are consistent with a neutral model of mutation and random drift. The spectrum of allele frequencies for amino acid variants shows a significant skew toward lower frequencies relative to both the neutral model and the pattern observed at linked silent sites. This is consistent with the hypothesis that weak purifying selection acted on TLR4 and that most mutations affecting TLR4 protein structure have at least mildly deleterious phenotypic effects. These results may imply that genetic variants contributing to disease susceptibility occur at low frequencies in the population and suggest strategies for optimizing the design of disease-mapping studies.

THE Toll-like receptor 4 protein (TLR4) acts as the transducing subunit of the lipopolysaccharide (LPS; endotoxin) receptor complex (Poltorak *et al.* 1998a,b; Du *et al.* 1999). Mutations of *Tlr4* are known to abolish responses to endotoxin in mice (Poltorak *et al.* 1998a; Hoshino *et al.* 1999), rendering these animals hypersusceptible to infection by Gram-negative bacteria (O'Brien *et al.* 1980; Rosenstreich *et al.* 1982; Yoshida *et al.* 1991; Macela *et al.* 1996) and insensitive to the toxic effects of lipopolysaccharide (Heppner and Weiss 1965; Coutinho *et al.* 1977; Coutinho and Meo 1978). On this basis, it has been inferred that TLR4 senses the presence of Gram-negative bacteria at an early stage of infection, permitting a timely response by the innate immune system. Generalizing from the examples provided by mutations of *Tlr4* in mice and *Toll* in Drosophila, loss-of-function mutations that affect Toll-like receptor structure or expression would be expected to impair host responses to a restricted spectrum of microbial pathogens. Arbour *et al.* (2000) have recently demonstrated that a common structural polymorphism of *TLR4* does, indeed, diminish airway responsiveness to aerosolized LPS in humans. Endotoxin exposure is associated with the development and progression of asthma and other forms of airway disease (Arbour *et al.* 2000). As such, it might be expected that genetic variability in mammalian Toll-like receptor genes could contribute to the pathogenesis of noninfectious as well as infectious diseases, particularly conditions that involve the inappropriate activation of mononuclear phagocytic cells.

Little is known about the pattern of *TLR4* sequence diversity in humans and how it might contribute to the susceptibility to infectious disease and sepsis. Studies of sequence variation may identify potential disease susceptibility variants to be tested in full-scale association studies, and they can also elucidate the evolutionary forces that acted on the gene. In particular, different patterns of variation are predicted by the different genetic models for susceptibility to complex traits. For example, one popular hypothesis (common disease-common variant; CD-CV) posits that common diseases with complex genetic and environmental etiology are due to co-inheritance of several variants that exist at high frequencies in the population, each contributing a small phenotypic effect. Alternatively, a much larger number of low-frequency variants may underlie disease susceptibility, as in the "multi-equivalent model" proposed by Wright *et al.* (1999). These competing models have implications for the efficiency and power of different disease-mapping strategies. Thus, elucidating the

Corresponding author: Bruce Beutler, The Scripps Research Institute, Department of Immunology (IMM-31), 10550 N. Torrey Pines Rd., La Jolla, CA 92037. E-mail: bruce@scripps.edu

[1]These authors contributed equally to this work.

mode of evolution of a gene may provide insights into the frequency distribution of variation contributing to disease susceptibility and assist in the optimization of the study design for disease mapping.

Here, we surveyed sequence variation in the entire *TLR4* coding region (2.4 kb) of 141 Caucasians, 45 African Americans, 25 Hispanic Americans, 48 individuals from Cameroon, and 89 individuals of an ethnically undefined population. In addition, 1.1 kb of the second intron of *TLR4* was sequenced in a subset of the same samples, *i.e.*, 50 Caucasians, and all the African Americans and Cameroonians. Our results show that there is a significant excess of low-frequency amino acid (aa) variants relative to the pattern observed for intronic and synonymous variants and to the expectations of a neutral equilibrium model. These results are consistent with a model of weak purifying selection, in which slightly deleterious variants rise to observable frequencies, but seldom go to fixation.

MATERIALS AND METHODS

DNA samples: The Cameroonian sample comprised 25 Hausa and 23 Beti from Yaounde. The Caucasian, African American, and Hispanic American samples were derived from an anonymized collection of DNA samples obtained from Dr. Ernest Beutler at The Scripps Research Institute, La Jolla, California. The ethnically undefined DNA samples were obtained from unselected ambulatory outpatient clinic patients at the University of Texas Southwestern Medical Center in Dallas, Texas.

PCR and sequencing: The *TLR4* gene, located at 9q32-q33, is comprised of three exons and spans ~10 kb. All three exons and a portion of intron 2 contiguous to exon 3 (positions 11075–12238 of accession no. AF177765) of the human *TLR4* locus were amplified using the primers shown in Table 1, and the products were either gel purified or separated from residual primers by means of spin dialysis over Sepharose CL4B. Both strands of the products were then directly sequenced using internal primers shown in Table 1. Fourteen reads were generally required to establish contiguous and partially overlapping high-quality sequence coverage on both strands throughout the *TLR4* coding region, and four reads were required for similar coverage of the intronic fragment. Dye terminator chemistry was used in these reactions, and sequences were resolved using ABI model 373 and 377 machines. The orthologous *TLR4* sequences were determined for a bonobo (*Pan paniscus*), a gorilla (*Gorilla gorilla*), an orangutan (*Pongo pygmaeus*), and a baboon (*Papio anubis*), which served as outgroups in the analysis.

Sequence analysis: Trace files obtained from each of the 348 human individuals and from each of the primate species were optimally assembled using the programs polyphred and Phrap (Nickerson *et al.* 1997). As a condition for further analysis, complete assembly (*i.e.*, generation of contigs corresponding to each of the three exons and, when applicable, the intronic sequence) was required. When necessary, additional reads were performed using a secondary set of primers to achieve assembly. Further analysis was performed by grouping all reads from 5 to 10 individuals, reassembling them, and examining them using the program consed (Gordon *et al.* 1998). Every base pair was inspected in every individual, with particular attention accorded to sites that were flagged by polyphred and sites that were found to be mutant in any one individual of the entire population. Phrap, polyphred, and consed were obtained from the University of Washington Genome Center and run on a DEC-alpha system (550 MHz, 256 Mbyte RAM) with a digital UNIX operating system. Every polymorphism occurring as a singleton in the total sample, except for the "Unknown," was confirmed by cloning and sequencing for each individual (nucleotide positions 11260, 11522, 11924, 13307, 13757, 14059, and 14478). Singletons occurring in the "Unknown" samples could not be confirmed because sufficient DNA was not available, and additional DNA could not be obtained due to prior agreement with the subjects participating in the study. The amplified fragments (either exon 3 or intron 2) were cloned into the vector pCR4 (Invitrogen, Carlsbad, CA; Topo TA cloning vector series). Six clones of each amplified product were selected and sequenced entirely on both strands using the same primers applied in direct sequencing of PCR fragments.

Data analysis: Sequences were analyzed by the program DnaSP 2.0 (Rozas and Rozas 1997) to obtain summary statis-

TABLE 1

Oligonucleotide primers used to amplify and sequence *TLR4*

	Amplification primers
EXON 1	F GCTCGGTAAACGGTGATAG
	R TGAGAAGTTCTGGGCAGAAG
EXON 2	F TCTCTGGTCTAGGAGAGG
	R CCAGTCCAATAATGAAATG
EXON 3	F CCATCACATCTGTATGAAGAGCTGGATGAC
	R TGACTTTCTTTGTCATGGGTTCCTTGACTG
	Sequencing primers
EXON 1	Same as for amplification
EXON 2	Same as for amplification
EXON 3	
Primary	3.1 F GAGCTGGATGACTAGGATTAATATTC
	3.1 R TCAAATTGCACAGGCCCTCTAG
	3.2 F CAATCTCTCTTTAGACCTGTCC
	3.2 R AATACTTTAGGCTGGTTGTCCC
	3.3 F GAAGTTGATCTACCAAGCCTTG
	3.3 R GGAAGTCATTATGTGATTGAGAC
	3.4 F CTTCCTGGACCTCTCTCAGTGTCAAC
	3.4 R GAAGGCAGAGCTGAAATGGAGG
	3.5 F TCAGATGAATAAGACCATCATTGGTG
	3.5 R AACAAGTGTTGGACCCAG
EXON 3	
Secondary	1 F GTAAATTTGGACAGTTTCC
	2 F TTCAGTATTCCTATCACTCAG
	3 F TTATAAGTGTCTGAACTCCC
	4 F TCGGTCCTCAGTGTGCTTG
	5 F GTGTCCCAGCACTTCATC
	6 F AACCTCCTGAGGCATTTC
	7 R GTTTCAAATTGGAATGCTG
	8 R AAGGAAACGTATCCAATG
	9 R AAGCACACTGAGGACCGAC
	10 R GATGAAGTGCTGGGACAC
	11 R TCCTCTTCAGATAGATGTTG
	12 R TTTCTTTGTCATGGGTTC
	0 F TTTAGGTTCTTATTCAGCAG
	0 R GCTCTAGATTGGTCAGATTAG

F, primer matches plus strand; R, primer matches minus strand.

tics of sequence variation and *D* (Tajima 1989). The significance of *D* was determined by coalescent simulations using a program written by J. D. Wall. Other statistical methods are cited in the text where appropriate.

RESULTS AND DISCUSSION

The locations of all polymorphic sites, and the frequencies of the nonancestral alleles, are shown in Table 2. For each polymorphic site, the nonancestral allele was determined by comparison to the same nucleotide position in the nonhuman primate sequences. Summary statistics of nucleotide diversity in these samples are shown in Table 3.

At the amino acid level, >90% of the sampled chromosomes carried the same *TLR4* allele. The vast majority of coding variants, largely defined by single amino acid changes, were present at low frequencies (1–7%) within the population groups in which they were observed and were found at still lower frequencies within the total human sample. A variant allele at a frequency of 7% (*TLR4B*; GenBank accession no. AF177766) was observed in Caucasians and was characterized by two amino acid substitutions in the ectodomain, located at residues 299 and 399 of the 839-aa polypeptide chain (variants 12874 and 13174 in Table 2).

Frequency spectra of silent and replacement variation: To evaluate the unusual spectrum of allele frequencies for amino acid polymorphisms, we employed the widely used statistic *D*. *D* is based on the difference between *k* (average sequence difference between all possible pairs of chromosomes) and θ_W (the number of polymorphic sites, corrected for sample size) and has an expectation near 0 for neutral variants at equilibrium (Tajima 1989). A significantly negative value indicates an excess of rare variants in the sample. As shown in Table 3, the amino acid variants have a significantly negative *D* value in the African American and pooled African samples as well as the total sample, indicating a departure from the neutral equilibrium model. Furthermore, sharply negative values are observed in all other population samples except the Caucasian. The assumption of panmixia in the neutral equilibrium model may be violated due to the structure of human populations, leading to the expectation of an increased variance of *D*. Additional sequence variation data from the same populations at unlinked loci will allow one to evaluate the relative roles of population structure and natural selection in shaping the frequency spectrum at this locus.

In contrast to the replacement polymorphisms, silent polymorphic sites (encompassing both intronic and synonymous positions) at *TLR4* show no significant skew toward lower-frequency variants. In fact, *D* for these sites is positive for the African American and Cameroonian samples (Table 3) and in agreement with a neutral equilibrium model.

The unusual pattern of allele frequencies at *TLR4* can be visualized by constructing a network of the amino acid variation only, based on inferred haplotypes (Figure 1). Because of the low frequency of heterozygotes at more than one site, the vast majority of the haplotypes could be unambiguously inferred. When two amino acid variants were observed in the same individual, they were inferred to be on the same chromosome, since the mutation-free chromosome is in such high frequency in all populations. In this network, two-thirds of the nonancestral haplotypes are one step removed from the most common haplotype and do not share mutations with other haplotypes. Only one haplotype is more than two steps removed from the most common haplotype. If the haplotype inference were incorrect, the network topology would be even more striking: mutations 13757, 14059, and 14478, for example, could each have occurred independently on the background of the most common haplotype.

An excess of low-frequency variants may result from a number of different evolutionary histories, some including natural selection, others including demographic changes, in particular population expansion (Tajima 1989). Demographic changes are expected to influence all sites in the nuclear genome in the same way, while selective forces tend to have short-range genomic effects. There are few examples of strongly negative *D* values at other nuclear loci sampled from human populations (Przeworski *et al.* 2000). This finding and, more importantly, the positive *D* values observed at silent sites at *TLR4* make population expansion an unlikely explanation for the excess of rare amino acid variants in our survey. The contrast between the spectrum of allele frequency at silent and at amino acid polymorphisms within the *TLR4* locus also makes a "selective sweep" (recent fixation of an advantageous mutation at or near *TLR4*) a less likely explanation for a significantly low *D*. Such an event would have affected silent as well as amino acid variation, and *all* sites in the region would be expected to show a skew toward rare alleles.

However, in this data set, the functional classification of silent and replacement polymorphisms also results in a spatial grouping: three of the four synonymous polymorphisms are at the 5′ end of exon 3 (physically linked to the intron, see Table 2), and all the amino acid polymorphisms are in exon 3. As a result, the difference in frequency spectra of intron and exon variants is similar to that of silent and replacement variants. This raises the possibility that a selective sweep occurred at or 3′ to exon 3, but failed to alter the frequency spectrum in the adjacent intron due to recombination between the two regions. The selective sweep model predicts a significant reduction of variation that can be assessed by means of the Hudson-Kreitman-Aguadé (HKA) test. This test uses divergence data to take account of differences in neutral mutation rates between loci (Hudson *et al.* 1987). Using orangutan as an out-

TABLE 2

Frequencies (counts) of the nonancestral alleles in the *TLR4* gene

Nucleotide position[a]	1	1	1	1	1	1	1	1	1	1						
	1	1	1	1	1	1	1	1	2	2						
	2	4	4	5	7	8	8	9	1	2						
	6	0	3	2	7	2	5	1	0	2						
	0	6	9	2	1	9	4	1	1	8						
							A. Intron 2 (1163 bp)									
Nonancestral allele	A	G	G	C	G	T	A	G	Δ	Δ						
African American (n = 90)	1	13	13	0	26	7	16	10	11	10						
Cameroonian (n = 96)	0	22	14	0	16	4	12	24	22	21						
Caucasian (n = 100)	0	4	0	1	75	0	22	0	0	0						
Total (n = 286)	1	39	27	1	117	11	50	34	33	31						
Nucleotide position[a]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4
	2	4	4	5	7	8	9	1	3	3	5	7	9	0	2	4
	9	1	6	4	1	7	6	7	0	9	0	5	3	5	6	7
	3	3	7	1	5	4	4	4	7	8	8	7	7	9	6	8
							B. Exon 3 (2240 bp)									
Nonancestral allele	T	A	G	G	C	G	G	T	A	A	T	A	A	G	A	A
Amino acid change	—	—	—	Q188R	C246S	D299G	N329S	T399I	F443L	E474K	Q510H	W593X	—	K694R	R763H	Q834K
African American (n = 90)	75	0	4	0	2	6	0	3	0	2	3	0	0	1	0	0
Cameroonian (n = 96)	77	0	0	0	0	4	0	0	0	5	4	1	0	0	0	1
Caucasian (n = 282)	282	0	0	0	0	17	0	17	1	0	0	0	3	0	0	0
Unknown (n = 178)	178	4	0	1	0	10	1	9	0	2	0	0	1	0	2	0
Hispanic (n = 50)	50	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Total (n = 696)	666	4	4	1	2	37	1	28	1	8	7	1	5	1	2	1

[a] With reference to GenBank accession no. AF177765.

TABLE 3

Summary statistics of variation at *TLR4*

	Nonsynonymous (1909 sites)				Intron and synonymous[a]			
Population	*n*	π	TD	FLD	*n*	π	TD	FLD
Cameroon	96	0.15	−1.48	−1.11	96	1.10	0.91	1.21
African American	90	0.17	−1.58*	−0.18	90	1.13	0.24	0.59
African pooled	186	0.16	−1.69*	−1.44	186	1.12	0.56	0.46
Caucasian	282	0.12	−0.78	−0.80	100	0.49	−0.27	−1.12
Unknown	178	0.15	−1.53	−0.99	178	0.10	−1.17	−1.21
Hispanic	50	0	–	–	50	0.07	−1.10	−1.87
Total	696	0.13	−1.82*	−3.45*	286	1.01	−0.03	−1.14

π is nucleotide diversity ×1000. TD is Tajima's (1989) *D*. FLD is Fu and Li's (1993) *D*. * *P* value < 0.05, using coalescent simulations without recombination.

[a] The entire coding region and 1.2 kb of intron 2 were resequenced in the Cameroonian, African American, and Caucasian samples; thus, the total number of surveyed intronic and synonymous sites was 1717. Only the coding region was resequenced in the Unknown and Hispanic samples; thus, only the 554 synonymous sites were included in this analysis for these samples. Total for intron and synonymous sites includes only the Cameroonians, African Americans, and 100 Caucasians.

group, we tested variation at TLR4 against variation in intron 44 of *DMD* (Nachman and Crowell 2000), one of the most variable loci in humans (Przeworski *et al.* 2000). The data were divided into exon and intron, rather than replacement and silent, because under a selective sweep model linked sites would be affected regardless of their functional classification. Using the same test, we also compared the TLR4 intron and exon variation to each other. None of these comparisons was significant (Table 4A). Variation in exon 3 of the Caucasian sample was low, and the test was close to significance, but since this sample does not show a significantly negative Tajima's *D*, this result does not bear on the question. The comparison of TLR4 intron to TLR4 exon variation, a direct test of the hypothesis that we have encountered the boundary of a selective sweep, has a particularly large *P* value (0.97), providing no support for this interpretation. This result is not surprising, given the tight linkage of these two regions. Furthermore, when we compared variation at TLR4 to variation at the β-globin locus, using bonobo as the outgroup (Table 4B), both the African and non-African samples show a marginal *excess* of polymorphism in exon 3. Thus, several aspects of the data appear to be inconsistent with the hypothesis that a selective sweep affected TLR4.

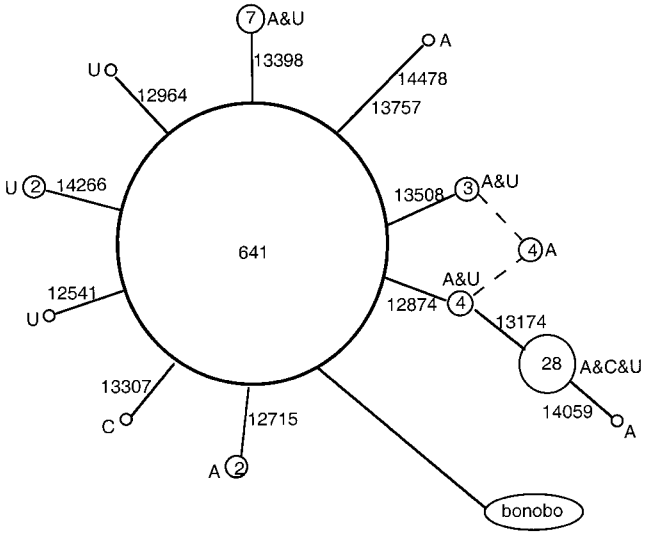

Figure 1.—Minimum-mutation network of amino acid variation at *TLR4*, based on inferred haplotypes. Each circle represents a haplotype; the number inside the circle indicates the number of occurrences of that haplotype. Unnumbered circles represent unique haplotypes. Letters next to circles indicate the population sample(s) in which the haplotype was observed. Numbers next to lines indicate the position of the mutation as in Table 2. Dotted lines indicate an inferred recombination event. A, African American and Cameroonian; C, Caucasian; U, unknown ethnicity. The circle containing the number 28 represents the *TLR4B* haplotype.

We also considered the possibility that the marked difference in *D* values between the intron and exon 3 of *TLR 4* was simply due to chance. To test this hypothesis, 10,000 coalescent simulations of the standard neutral model with recombination were carried out as follows: gene genealogies were generated for a 3.4-kb region, which was subsequently divided into two regions of 1.2 kb and 2.2 kb corresponding to the TLR4 intron and exon 3, respectively. A difference in *D* values as large as or larger than that observed for intron and exon 3 variants in the pooled African sample was found in only 7% of the simulations. Although this percentage is only marginally significant, it should be pointed out that the simulations did not include the condition that one of the *D* values be significantly negative, as observed in this study. Thus, the standard neutral model does not provide an adequate description of our data, and this analysis may be overly conservative.

Having found no support for demographic explana-

TABLE 4

HKA tests of neutrality

A. Orangutan as an outgroup

	Tests within *TLR4*				Tests *vs.* DMD intron 44					
	African		Caucasian		African			Caucasian		
	TLR4 exon 3	*TLR4* intron	*TLR4* exon 3	*TLR4* intron	*TLR4* exon 3	*TLR4* intron	DMD	*TLR4* exon 3	*TLR4* intron	DMD
n	186	186	282	100	186	186	10	282	100	10
S	10	7	4	4	10	7	14	4	4	10
l	2240	1063	2240	1063	2240	1063	3000	2240	1063	3000
D	39	28	39	28	39	28	78	39	28	78
P	0.97		0.52		0.31	0.63		0.08	0.32	

B. Bonobo as an outgroup

	African		Non-African	
	β-Globin	*TLR4* exon 3	β-Globin	*TLR4* exon 3
n	206	186	1048	510
S	16	10	15	9
l	2320	2240	2320	2240
D	27	5	27	5
P		0.08		0.06

n, number of chromosomes; S, number of segregating sites; l, number of bases surveyed; *D*, average pairwise divergence to outgroup; *P*, *P* value.

tions or a selective sweep, weak purifying selection on the rare amino acid variants in *TLR4* remains as a viable explanation. According to population genetics theory (Kimura 1983), mutations for which $N_e s < 1$ (where $N_e s$ is the product of selection coefficient and effective population size) will behave as "nearly neutral," and their behavior will be largely a function of genetic drift.

Long-standing diversifying selection would lead to an elevated level of polymorphism and a positive *D*, neither of which we observed. Another possibility is that the amino acid variants segregating at low frequency today might be under very recent diversifying selection. This possibility cannot be excluded and is weakly supported by the marginal excess of variation in the comparison to the β-globin gene (Table 4B). However, recent diversifying selection cannot be easily reconciled with the fact that human populations have been exposed to Gram-negative pathogens throughout their evolutionary history.

Interspecific comparisons: For strictly neutral mutations, the ratio of amino acid to synonymous variants within species is expected to be the same as that observed between species. However, because slightly deleterious mutations tend to be eliminated before they reach high frequencies, they are more likely to be observed among within-species variants than among fixed substitutions between species. We compared synonymous and amino acid polymorphisms within humans to fixed differences between human and bonobo, gorilla, orangutan, and baboon. This analysis did not include sequence data from the intron. A departure from the neutral model was not observed for any comparison (Table 5). The failure to detect a significant excess of

TABLE 5

Tests of neutral evolution based on polymorphism and divergence

	S	R
Human to		
Bonobo	3	3
Gorilla	6	5
Orangutan	17.5	21.5
Baboon	31	54
Within populations		
African American	2	6
Cameroon	1	5
Caucasian	1	3
Unknown	2	6
Total	4	12

S and R indicate the number of synonymous and amino acid replacement changes, respectively. The test of McDonald and Kreitman (1991) was implemented using Fisher's exact test.

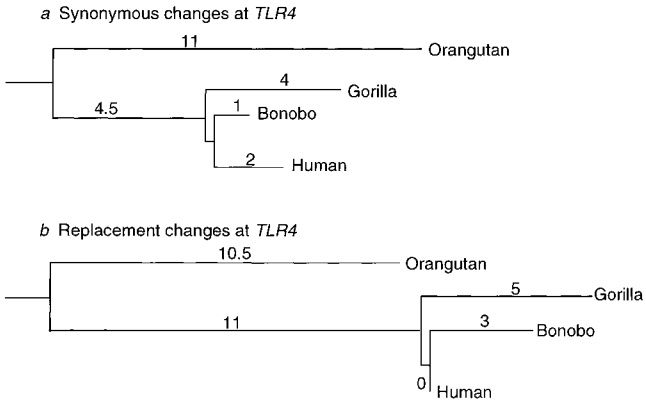

FIGURE 2.—Rates of synonymous and amino acid evolution at *TLR4* in primates. Branch lengths are proportional to the number of changes as estimated by the method of SARICH and WILSON (1973). The consensus phylogeny of these species was assumed (GOODMAN *et al.* 1998). The baboon lineage is not represented because it was used as the outgroup for the relative rate test.

amino acid polymorphisms relative to divergence from the outgroup is likely the result of the low power of the test. Another possibility is that evolutionary rates differ across different lineages of the primate phylogeny. Protein evolution along the branch from the common ancestor of orangutan and human to the common ancestor of bonobo and human appears to have been faster relative to silent changes and to protein evolution in the human and bonobo lineages (Figure 2). This pattern suggests that a change in constraints in the human and bonobo lineages might underlie the apparent contradiction between a significant excess of rare variants and no excess of amino acid polymorphisms. Thus, these interspecific comparisons are consistent with the hypothesis that weak purifying selection is the major evolutionary force acting on protein level evolution at *TLR4* in the human lineage.

Slightly deleterious amino acid variation: An implication of the proposal that weak purifying selection acted on *TLR4* is that a portion of the amino acid variants observed have phenotypic effects that reduce the fitness. Gram-negative pathogens such as *Yersinia pestis*, *Salmonella typhi*, *Rickettsia prowazekii*, and *Neisseria meningitidis* have exerted strong selective pressures on populations within recorded human history, and these and other agents may have done so in the remote past as well. Mutations that diminish the ability of TLR4 protein to detect pathogens would certainly be disfavored in the population and might at most achieve modest frequencies, perhaps during intervals of time when no selective agent is prevalent. However, TLR4 fulfills a delicate and somewhat dangerous role in the mammalian host. Although represented in small numbers on the surface of mononuclear cells (DU *et al.* 1999), TLR4 delivers a potentially lethal pro-inflammatory signal. Mutations of TLR4 might confer hypersensitivity to LPS or cause constitutive signaling activity via the LPS receptor, either of which would clearly be deleterious. Only truly neutral mutations, which exert no effect on the sensing function of TLR4 and do not result in constitutive signaling, are likely to be retained within the population over the long term.

Slightly deleterious mutations may be a common feature of human genome diversity. In line with this idea, SUNYAEV *et al.* (2000) have recently reported an excess of rare amino acid variants in genome-wide surveys of coding sequence variation, suggesting that a subset of these variants is slightly deleterious. It should be noted that the detection of this phenomenon at a single locus, such as *TLR4*, requires that a large number of nonsynonymous sites be surveyed in large population samples. Our study design was appropriate for this goal, representing the single largest coding region survey in a very large sample. Furthermore, the analysis of intronic sequence variation in a large subset of the same samples allowed us to contrast the pattern of variation at nonsynonymous and silent sites with greater power and, therefore, rule out alternative demographic and adaptive explanations. Evidence for a slightly deleterious mutation model applying to an individual locus was previously reported at a candidate gene for cardiovascular disease, the lipoprotein lipase gene (*LPL*), which shows a significant excess of amino acid changes within, relative to between, species (CLARK *et al.* 1998; NICKERSON *et al.* 1998). Interestingly, one of these amino acid variants is associated with premature atherosclerosis (REYMER *et al.* 1995). We used the same analytical approach on the *LPL* data set and determined that, as for *TLR4*, *D* for amino acid polymorphisms is significantly negative (-1.73, $P < 0.05$) while it is positive for the silent ones.

This study suggests that, while this mode of evolution might affect a portion of all coding variants in the human genome, it can also generate a significant skew in the pattern of variation of an individual gene. Although the theoretical framework for "nearly neutral" evolution is still debated (OHTA and GILLESPIE 1996), this study contributes to the mounting empirical evidence supporting the existence of this class of mutations. It follows that coding sequence variation data are not suitable for inferring population histories since, as shown here, weak purifying selection may generate a multi-locus pattern of allele frequencies that is not related to human demography.

Implications for disease mapping: If many coding variants occur at low frequency and have deleterious phenotypic effects, it may be postulated that rare mutations play a larger role in common diseases than is often assumed in disease-mapping strategies. The greater allelic heterogeneity would translate into a major challenge for disease association studies, especially in outbred populations. It has been argued that, if a multitude of rare variants (rather than a restricted number of common variants) underlie the genetic susceptibility

to common diseases, linkage mapping strategies would prove more powerful than linkage disequilibrium-based mapping. Moreover, because founder events reduce the number of rare alleles but have little effect on common variants, recent founder populations would have lower allelic heterogeneity specifically with regard to slightly deleterious mutations (Wright *et al.* 1999). Thus, if the mode of evolution of a candidate gene appears to fit a slightly deleterious mutation model, as is the case for *TLR4* and *LPL*, then linkage rather than linkage disequilibrium mapping strategies, applied to recent founder rather than outbred populations, might be a more productive approach.

We are grateful to A. Pluzhnikov for carrying out coalescent simulations. We thank B. Charlesworth, A. Clark, R. Hudson, C. Ober, and A. Turkewitz for helpful discussions and comments on the manuscript. We thank J. Donfack for DNA samples. We thank J. D. Wall for determining the significance of *D.* M.H. and A.D. were partially supported by National Institutes of Health (NIH) grant R01-HG02098 to A.D.; I.S., C.M., and B.B. were supported by NIH grant 1-R01-GM60031-01 and by the Howard Hughes Medical Institute.

LITERATURE CITED

Arbour, N. C., E. Lorenz, B. C. Schutte, J. Zabner, J. N. Kline *et al.*, 2000 *TLR4* mutations are associated with endotoxin hyporesponsiveness in humans. Nat. Genet. **25:** 187–192.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Coutinho, A., and T. Meo, 1978 Genetic basis for unresponsiveness to lipopolysaccharide in C57BL/10Cr mice. Immunogenetics **7:** 17–24.

Coutinho, A., L. Forni, F. Melchers and T. Watanabe, 1977 Genetic defect in responsiveness to the B cell mitogen lipopolysaccharide. Eur. J. Immunol. **7:** 325–328.

Du, X., A. Poltorak, M. Silva and B. Beutler, 1999 Analysis of Tlr4-mediated LPS signal transduction in macrophages by mutational modification of the receptor. Blood Cells Mol. Dis. **25:** 328–338.

Fu, Y.-X., and W.-H., Li 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider *et al.*, 1998 Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9:** 585–598.

Gordon, D., C. Abajian and P. Green, 1998 Consed: a graphical tool for sequence finishing. Genome Res. **8:** 195–202.

Heppner, G., and D. W. Weiss, 1965 High susceptibility of strain A mice to endotoxin and endotoxin-red blood cell mixtures. J. Bacteriol. **90:** 696–703.

Hoshino, K., O. Takeuchi, T. Kawai, H. Sanjo, T. Ogawa *et al.*, 1999 Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J. Immunol. **162:** 3749–3752.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, United Kingdom.

Macela, A., J. Stulik, L. Hernychova, M. Kroca, Z. Krocova *et al.*, 1996 The immune response against *Francisella tularensis* live vaccine strain in Lps^n and Lps^d mice. FEMS Immunol. Med. Microbiol. **13:** 235–238.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

Nachman, M. W., and S. L. Crowell, 2000 Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, DMD, in humans. Genetics **155:** 1855–1864.

Nickerson, D. A., V. O. Tobe and S. L. Taylor, 1997 PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. **25:** 2745–2751.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.*, 1998 DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. **19:** 233–240.

O'Brien, A. D., D. L. Rosenstreich, I. Scher, G. H. Campbell, R. P. Macdermott *et al.*, 1980 Genetic control of susceptibility to Salmonella typhimurium in mice: role of the LPS gene. J. Immunol. **124:** 20–24.

Ohta, T., and J. H. Gillespie, 1996 Development of neutral and nearly neutral theories. Theor. Popul. Biol. **49:** 128–142.

Poltorak, A., X. He, I. Smirnova, M.-Y. Liu, C. Van Huffel *et al.*, 1998a Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. Science **282:** 2085–2088.

Poltorak, A., I. Smirnova, X. L. He, M. Y. Liu, C. Van Huffel *et al.*, 1998b Genetic and physical mapping of the *Lps* locus-identification of the toll-4 receptor as a candidate gene in the critical region. Blood Cells Mol. Dis. **24:** 340–355.

Przeworski, M., R. R. Hudson and A. DiRienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Reymer, P. W., E. Gagne, B. E. Groenemeyer, H. Zhang, I. Forsyth *et al.*, 1995 A lipoprotein lipase mutation (Asn291Ser) is associated with reduced HDL cholesterol levels in premature atherosclerosis. Nat. Genet. **10:** 28–34.

Rosenstreich, D. L., A. C. Weinblatt and A. D. O'Brien, 1982 Genetic control of resistance to infection in mice. CRC Crit. Rev. Immunol. **3:** 263–330.

Rozas, J., and R. Rozas, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

Sarich, V. M., and A. C. Wilson, 1973 Generation time and genomic evolution in primates. Science **179:** 1144–1147.

Sunyaev, S., W. C. Lathe, V. E. Ramensky and P. Bork, 2000 SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.

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

Wright, A. F., A. D. Carothers and M. Pirastu, 1999 Population choice in mapping genes for complex diseases. Nat. Genet. **23:** 397–404.

Yoshida, S., Y. Goto, H. Miyamoto, H. Fujio and Y. Mizuguchi, 1991 Association of Lps gene with natural resistance of mouse macrophages against Legionella pneumophila. FEMS Microbiol. Immunol. **4:** 51–56.

Communicating editor: D. Charlesworth