Global Footprints of Purifying Selection on Toll-Like Receptor Genes Primarily Associated with Response to Bacterial Infections in Humans

Souvik Mukherjee[1,2,*], Debdutta Ganguli[1], and Partha P. Majumder[1]

[1]National Institute of Biomedical Genomics, Kalyani, West Bengal, India

[2]Present address: Biomedical Genomics Centre, Kolkata, West Bengal, India

*Corresponding author: E-mail: sm2.bmgc@nibmg.ac.in, callsouvik@gmail.com.

Accepted: February 12, 2014

Abstract

Toll-like receptors (TLRs) are directly involved in host–pathogen interactions. Polymorphisms in these genes are associated with susceptibility to infectious diseases. To understand the influence of environment and pathogen diversity on the evolution of TLR genes, we have undertaken a large-scale population-genetic study. Our study included two hunter–gatherer tribal populations and one urbanized nontribal population from India with distinct ethnicities ($n = 266$) and 14 populations inhabiting four different continents ($n = 1,092$). From the data on DNA sequences of cell-surface TLR genes, we observed an excess of rare variants and a large number of low frequency haplotypes in each gene. Nonsynonymous changes were few in every population and the commonly used statistical tests for detecting natural selection provided evidence of purifying selection. The evidence of purifying selection acting on the cell-surface TLRs of the innate immune system is not consistent with Haldane's theory of coevolution of immunity genes, at least of innate immunity genes, with pathogens. Our study provides evidence that genes of the cell-surface TLRs, that is, *TLR2* and *TLR4*, have been so optimized to defend the host against microbial infections that new mutations in these genes are quickly eliminated.

Key words: innate immune system, Toll-like receptors, purifying selection, evolution, haplotype, Indian populations.

Introduction

The innate immune system is ancient; it is the first line of host defense against invading pathogens (Kimbrell and Beutler 2001). In innate immunity, microbial recognition is mediated by a set of cell-surface pattern recognition receptors (PRRs) that recognize pathogen-associated molecular patterns (PAMPs) present in bacteria and other pathogens, but absent in higher eukaryotes (Texereau et al. 2005). Views on the mode of evolution of genes of the innate immune system are opposing: 1) natural selection has functionally optimized these genes so that newly arising mutations are quickly lost from a population; or 2) these genes co-evolve with rapidly evolving microbes (Parham 2003; drawing support from Haldane 1949). Both views have found support in empirical studies. Purifying selection identified in some studies (Barreiro et al. 2009; Mukherjee et al. 2009; Seabury et al. 2010) has supported the first view, while some other studies have identified balancing or positive selection (Ferrer-Admetlla et al. 2008; Wlasiuk and Nachman 2010; Areal et al. 2011) consistent with the alternate view. These differences in inference probably arose because data from restricted geographical regions or ethnicities were analyzed. The present was undertaken to overcome this major limitation, because to understand evolutionary patterns and modes from data on contemporary populations it is imperative to ensure a wide geographical and ethnic coverage.

As agents of natural selection, infectious diseases have played a major role in the evolution of human species. The advent of agriculture in human societies is believed to have been a significant event in the evolution of human diseases within the past 11,000 years (Wolfe et al. 2007). Organized agriculture and its geographical spread resulted in increase of population densities and human movements that permitted the persistence and spread of "civilization pathogens," such as measles, small pox, malaria, and so forth, which transmit rapidly from one person to another, often resulting in death of the infected person (Fiennes 1978; McNeill 1989; Diamond 1997, 2002). The appearance and transmission of infectious pathogens in human societies is dependent on the environment. Thus, there are wide geographical differences in the nature and prevalence of infectious diseases in human populations. Many emerging infectious diseases are associated with human modification of the environment (Epstein 1995; Schrag and Wiener 1995; Daszak et al. 2000).

We have, therefore, chosen our study populations from diverse environmental habitats and with distinct ancestries. We have performed molecular evolutionary analysis of data on genes from a large number of ethnically diverse populations from all four major continents. Two genes of the innate immune system, *TLR2* and *TLR4*, have been most extensively characterized and are the major effectors for both the Gram-negative (*TLR4*) and Gram-positive (*TLR2*) bacterial ligands (Takeuchi et al. 1999; Elson et al. 2007). These two genes can, therefore, serve as models to study how infectious agents may have shaped our cell-surface Toll-like receptors (TLRs) in particular and our innate immune system in general.

Materials and Methods

Study Populations and DNA Resequencing

To capture disparate selection pressures and regimes, we have chosen our study populations from diverse environmental habitats and with distinct ancestries. Unrelated individuals from two hunter-gatherer populations in India, an Austro-asiatic tribe, Bison Horn Maria ($n = 47$), from Madhya Pradesh, and a Dravidian tribe, Irula ($n = 48$), from Tamil Nadu, and an Indo-European caste population ($n = 171$) inhabiting an urban area of West Bengal, India, were included in this study.

A sample of blood was collected from each study participant by venipuncture with written informed consent, after obtaining institutional ethical approval for the study. DNA was isolated from the blood samples either by the salting out method (Miller et al. 1988) or by Qiagen-Maxi column DNA Isolation kit.

Double-pass resequencing of the exons, exon–intron boundaries and approximately 2 kb of each of 5′-upstream and 3′-downstream of *TLR2* and *TLR4*, using a capillary sequencer (ABI-3730), was done, with standard quality checks. Approximately 16 kb of the genome was resequenced for each individual (supplementary table S1, Supplementary Material online).

Analyses of sequence chromatograms and genotype calls were carried out using SeqScape v2.5 (from Applied Biosystems) and PolyPhred (http://droog.gs.washington.edu/polyphred/, last accessed March 3, 2014) software packages. Coding regions were translated using DNASTAR and BioEdit packages.

Data from Public-Domain Databases

To include data encompassing global diversity of human habitats and ancestries, we have augmented the data from India with those available in the 1000 Genomes database (http://www.1000genomes.org/, last accessed March 3, 2014) (The 1000 Genomes Project Consortium 2012). Brief details of the populations, with sample sizes, are provided in table 1.

Table 1
Description of the Study Populations with Their Sample Sizes

Population ID	Population Name	Sample Size
Asia		
IND[a]	Urban population from Kolkata, India	171
MAR[a]	Bison Horn Maria from Madhya Pradesh, India	47
ILA[a]	Irula Tribe from Tamil Nadu, India	48
CHB	Han Chinese in Beijing, China	97
CHS	Southern Han Chinese	100
JPT	Japanese in Tokyo, Japan	89
Africa		
ASW	African American in Southwest US	61
LWK	Luhya in Webuye, Kenya	97
YRI	Yoruba in Ibadan, Nigeria	88
America		
CLM	Colombian in Medellin, Colombia	60
MXL	Mexican-American in Los Angeles, California	66
PUR	Puerto Rican in Puerto Rico	55
Europe		
CEU	Utah residents with Northern and Western European ancestry	87
FIN	Finnish in Finland	93
GBR	British in England and Scotland	89
IBS	Iberian populations in Spain	14
TSI	Tuscan in Italia	98

[a]DNA sequence data on this population was generated in this study.

Variant Calling Format (VCF) files were extracted from the 1000 Genomes database for both the *TLR2* and *TLR4* genes using Data Slicer (http://browser.1000genomes.org/Homo_sapiens/Search/, last accessed March 3, 2014). Genotype files were generated from VCF files using computer programs developed by us. Reference Sequences of the chimpanzee, gibbon, and rhesus were downloaded for the two *TLR* genes from the UCSC genome browser.

Statistical Analyses

Estimation of allele frequencies and tests of Hardy–Weinberg equilibrium (HWE) were carried out using PLINK (Purcell et al. 2007). The average observed heterozygosities were estimated considering the total number of variant sites across all the 14 populations present in the VCF file for each gene (166 loci for *TLR4* and 322 loci for *TLR2*), downloaded from the 1000 Genomes website. For the Indian populations, the total number of variant loci across all the three populations was considered for estimating the average heterozygosity values for both the genes (53 loci for *TLR4* and 28 loci for *TLR2*).

Pairwise F_{ST} (Fixation Index) values for the common single nucleotide variants (SNVs) in both the genes (no. of SNVs = 30) among the 14 continental populations were

Table 2
Number and Nature of Variants in *TLR2* and *TLR4* Genes in Three Indian Populations

Gene	Population Name	Total Number of Variants	No. of Polymorphic Variants	No. of Nonpolymorphic Variants	No. of Coding Variants	No. of Synonymous Variants	No. of Nonsynonymous Variants	No. of Variants Shared by All Three Populations
TLR2	Tribe							
	Bison Horn Maria ($n = 47$)	8	4	4	3	3	0	6
	Irula ($n = 48$)	7	4	3	2	2	0	
	Caste							
	Urban Indian ($n = 171$)	25	6	19	8	5	3	
TLR4	Tribe							
	Bison Horn Maria ($n = 47$)	20	14	6	3	0	3	17
	Irula ($n = 48$)	23	16	7	5	1	4	
	Caste							
	Urban Indian ($n = 171$)	46	17	29	9	3	6	

estimated using Arlequin v3.5 (Excoffier and Lischer 2010). The data for three Indian populations were not included in the F_{ST} estimation, since they shared very few (9) SNVs with the continental populations for these genes. Haplotype identification and estimation of haplotype frequencies were done using PHASE for Windows, version 2.1 (http://stephenslab.uchicago.edu/software.html#phase, last accessed March 3, 2014) (Stephens et al. 2001; Stephens and Donnelly 2003). Statistics for evaluating departures of allele frequency spectra from neutrality (Tajima's *D*, Fu and Li's *D** and *F**, and Fu's F_s values) were computed using DnaSP, version 4.10 (Rozas et al 2003). Coalescent simulations, using DnaSP were carried out to test the statistical significance of Fu's F_s. DnaSP was also used for calculation of haplotype diversity. To assess the nature and extent of natural selection on these genes, we estimated the rates of nonsynonymous (d*N*) and synonymous (d*S*) substitutions by the Nei–Gojobori method (Nei and Gojobori 1986) using PAML4 (Yang 2007). Because different haplotypes were represented in varying numbers of individuals for each gene, we took care to represent each distinct haplotype by its observed frequency in the input file of PAML4 used to obtain the estimates of d*N* and d*S*. Weighted averages of resulting d*N*:d*S* estimates were taken.

In our study, for each gene the extended homozygosity of the most frequent haplotype was calculated around a "core" using the method suggested by Sabeti et al. (2002). The "core" was taken to be a centrally located single nucleotide polymorphisms (SNP) within the gene and with low heterozygosity. Data on a random sample of 5000 of 25,000 simulated haplotypes generated under the neutral model, using the computer program "ms" by Hudson (2002) with $\theta = 1$, were analyzed for extended haplotype homozygosity ($\theta = 0.5$ did not produce any strikingly different result). Haplotype networks were drawn for both the genes by the Median joining method, using the Phylogenetic Network Software NETWORK 4.2.0.1, website: fluxus-engineering.com (Bandelt et al. 1999). DNA sequence conservation was assessed by alignment with the three orthologous primate reference sequences, viz., chimpanzee, gibbon, and rhesus.

Results

Pan-India Variation in *TLR2* and *TLR4*

In the urban Indian nontribal population, a total of 71 variants were identified in the two genes, of which 23 were polymorphic (minor allele frequency > 0.05). Among the 17 coding variants identified, 9 were found to be nonsynonymous (table 2). We discovered 44 novel variants in these two genes that were previously unreported; these data were submitted to dbSNP and rs ids have been assigned (supplementary table S2, Supplementary Material online). In the two tribal populations—Bison Horn Maria (MAR) and Irula (ILA)—we identified a total of 58 variants, of which 38 were polymorphic (table 2), 8 (~14%) were novel, that is, unreported in dbSNP (supplementary table S3, Supplementary Material online), and 13 (~22%) were in the coding region, 7 of which were nonsynonymous. All SNPs were in HWE in all the three populations. The proportions of polymorphic variants were found to be higher in the rural tribal populations than in the urban population (table 2). Interestingly, however, the tribal populations harbored fewer haplotypes for each gene than the urban population (data not shown). These observations may be a reflection of higher antiquity, but greater isolation and smaller population size, of tribals than castes.

Twenty-three variant loci (6 in *TLR2* and 17 in *TLR4*) were shared by all the three populations; allele frequency differences among the populations at each of the 20 loci (supplementary table S4, Supplementary Material online) were statistically nonsignificant ($P > 0.05$). The three SNPs that showed significant differences in allele frequencies across

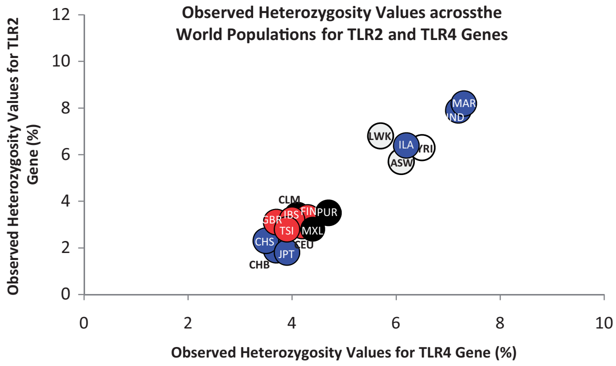

Fig. 1.—Observed heterozygosity values across World populations for *TLR2* and *TLR4* genes.

the Indian populations were rs4696480, rs1898830 of *TLR2* and rs11536889 of *TLR4* ($P < 0.05$). These SNPs are also found to be associated with regulation of gene expression and disease susceptibility in different studies (Budulac et al. 2012; Ito et al. 2012; Nischalke et al. 2012; Sato et al. 2012) and hence may have an important relevance in the Indian context where infectious diseases continue to be a major threat to human health. The average heterozygosity values of the two genes in the populations were plotted as a scatter. Figure 1 indicates 1) low gene diversity in each of the genes in all the populations, and 2) gene diversities of the two genes are correlated.

Transcontinental Variation in *TLR2* and *TLR4*

The number of SNVs is highly variable across the 14 global populations included in the 1000 Genomes Project, with populations of predominantly African ancestry harboring the highest number of SNVs (e.g., 160 variants in *TLR2* gene for LWK). Approximately 40% of the variant sites were polymorphic (Minor Allele Frequency >0.05) in each gene (supplementary table S5, Supplementary Material online). The average heterozygosity for each of the two genes was low and ranged from 6% in African populations (YRI, ASW, and LWK) to 2–3% in European and Asian populations (fig. 1). The number of haplotypes was, however, high in each population, particularly in the populations (ASW, LWK, and YRI) with African ancestry. Low frequency haplotypes were mostly present in these three populations. For each gene, three or four haplotypes were in high frequencies in all populations. Populations sharing ancestry also shared haplotypes (supplementary tables S6 and S7, Supplementary Material online). Haplotype diversities were very high (>90%) in all the populations (supplementary table S8, Supplementary Material online). The large number of haplotypes with high haplotype diversities indicates the presence of excess of rare variants (minor allele frequency < 0.05) across populations. Presence of a small number of modal haplotypes also suggests that natural selection may be the major determinant in selecting a small number of variants to rise in frequency while disfavoring others.

Pairwise F_{ST} Values Are Low among Continental Populations Except for Africa

F_{ST} (Fixation Index) is a measure of genetic differentiation among populations and is based on expected heterozygosities of the total population (H_T) relative to the subpopulations (H_S). Estimates of F_{ST} based on data on loci from different genomic regions provide valuable insights on selection operating in different regions of the genome if the variable demographic histories of populations are taken into account. Recent studies have revealed that F_{ST} among world populations is 0.05 for microsatellite data and 0.10 for the SNP data (Rosenberg et al. 2002; Li et al. 2008; Holsinger and Weir 2009).

Contrary to the previous reports, except for the African populations for whom the F_{ST} values are high (mostly ranging within 0.10–0.20) when compared with populations from Europe, Asia, or America, the values are quite similar (mostly less than 0.05) among populations of other continents (supplementary table S9, Supplementary Material online). The similar and small F_{ST} values among the continental populations are again consistent with a dominant role of natural selection on these two genes.

Statistical Evidence of Purifying Selection on *TLR2* and *TLR4* Genes by Analysis of Data on Global Populations

Statistical tests of neutrality (Tajima's *D*, Fu Li's *D**, *F**, and Fu's F_s) were performed on the genotype data separately for *TLR2* and *TLR4* and separately for all the 17 populations. For both the genes, in most of the populations, estimates of these statistics were negative (table 3), even though the estimates were not significantly different from zero. The consistent pattern of negative estimates for all the four relevant statistics in a population for each gene is indicative of purifying selection having operated on these genes. This may also be due to selective sweeps or population expansion, although this is not supported by results of further analyses presented later.

The ratio of nonsynonymous to synonymous changes are less than unity ($dN/dS < 1$) for both the genes in all the study populations (table 4) when compared with three primate ancestral sequences (chimpanzee, rhesus, and gibbon). This is a strong evidence of conservation of the coding region, indicating that new variations are not tolerated. Taken together with the consistently negative estimates obtained in respect of the neutrality statistics, the evidence of purifying selection operating on these TLRs is overwhelmingly strong.

Extended Haplotype Homozygosity Patterns Do Not Show Signs of Recent Positive Selection

If a recent mutation increases to a high frequency in a short period of time due to positive selection, then most individuals will be homozygous at the polymorphic loci spanning a large

Table 3

Observed Number of Haplotypes and Estimates of Haplotype Diversity and Statistics for Testing Neutrality by Population Group for *TLR2* and *TLR4* Genes

Population Code	*TLR2*						*TLR4*					
	No. of Haplotypes	Haplotype Diversity	Tests of Selective Neutrality				No. of Haplotypes	Haplotype Diversity	Tests of Selective Neutrality			
			Tajima's *D*	FuLi's *D**	FuLi's *F**	Fu's *Fs*			Tajima's *D*	FuLi's *D**	FuLi's *F**	Fu's *Fs*
ASW	60	0.977	−0.95	−0.58	−0.89	−10.58^{a}	42	0.958	−0.90	−0.05	−0.49	−10.27^{b}
CEU	47	0.912	0.05	−1.72	−1.12	−4.62	24	0.865	−0.77	−0.098	−0.46	−2.61
CHB	46	0.902	−0.13	−3.19^{a}	−2.19^{c}	−5.77	25	0.837	−0.98	−1.29	−1.40	−7.31^{a}
CHS	49	0.863	0.47	−1.02	−0.44	−8.00	27	0.849	−0.78	−1.61	−1.53	−7.34^{a}
CLM	43	0.908	−0.63	−2.38^{a}	−1.95^{c}	−5.20	31	0.889	−1.48	−0.32	−0.96	−8.05^{a}
FIN	47	0.918	0.58	1.32	1.19	−4.16	19	0.873	−0.0001	0.94	0.66	0.63
GBR	51	0.906	−0.09	−0.99	−0.70	−7.40	25	0.842	−0.91	−0.004	−0.47	−3.22
IBS	13	0.894	0.26	0.98	0.88	2.09	10	0.878	0.27	0.53	0.53	0.01
JPT	35	0.917	0.59	−0.94	−0.37	−3.37	24	0.793	−0.75	−1.25	−1.26	−7.05
LWK	65	0.943	−0.96	0.16	−0.45	−7.48	71	0.971	−0.82	−0.46	−0.75	−30.97^{b}
MXL	51	0.954	−0.58	−2.36^{a}	−1.89^{c}	−9.30^{a}	30	0.887	−1.60^{c}	−2.53^{a}	−2.57^{a}	−9.32^{b}
PUR	46	0.932	−1.04	−4.09^{b}	−3.32^{b}	−6.41	26	0.89	−1.50	0.47	−0.41	−5.03
TSI	59	0.923	−0.01	−1.51	−0.99	−11.11^{a}	28	0.853	−1.08	−0.55	−0.94	−5.35
YRI	70	0.965	−0.73	0.98	0.21	−11.55	58	0.968	−0.29	−0.095	−0.22	−17.79
IND	34	0.659	−1.24	−2.46^{a}	−2.35^{a}	−22.34^{a}	58	0.78	−1.23	−3.50^{b}	−2.98^{b}	−34.49^{a}
MAR	9	0.576	0.34	1.26	1.12	−0.68	15	0.827	0.18	1.30	1.05	−0.50
ILA	6	0.715	0.61	0.33	0.50	1.41	18	0.834	0.61	−0.19	−0.40	−2.99

[a] $P < 0.05$.
[b] $P < 0.02$.
[c] $0.10 > P > 0.05$.

Table 4

Rates of Nonsynonymous (d*N*) and Synonymous (d*S*) Substitutions Per Site and Their Ratios for the *TLR2* and *TLR4* Gene Compared with Ancestral Primate Sequences

Population Code	Coding Region of *TLR2* Gene Compared with Ancestral Sequences of									Coding Region of *TLR4* Gene Compared with Ancestral Sequences of								
	Rhesus			Gibbon			Chimpanzee			Rhesus			Gibbon			Chimpanzee		
	d*N*	d*S*	d*N*/d*S*	d*N*	d*S*	d*N*/d*S*	d*N*	d*S*	d*N*/d*S*	d*N*	d*S*	d*N*/d*S*	d*N*	d*S*	d*N*/d*S*	d*N*	d*S*	d*N*/d*S*
ASW	0.019	0.063	0.298	0.015	0.045	0.332	0.002	0.010	0.215	0.027	0.059	0.466	0.011	0.025	0.444	0.002	0.005	0.435
CEU	0.019	0.063	0.298	0.015	0.046	0.330	0.002	0.011	0.211	0.027	0.059	0.464	0.011	0.025	0.435	0.002	0.005	0.398
CHB	0.019	0.063	0.297	0.015	0.046	0.324	0.002	0.011	0.193	0.027	0.059	0.463	0.011	0.025	0.436	0.002	0.005	0.395
MXL	0.019	0.063	0.297	0.015	0.046	0.326	0.002	0.011	0.201	0.027	0.059	0.464	0.011	0.025	0.435	0.002	0.005	0.396
YRI	0.019	0.063	0.297	0.015	0.045	0.333	0.002	0.010	0.217	0.027	0.059	0.466	0.011	0.025	0.445	0.002	0.005	0.439
IND	0.019	0.062	0.299	0.015	0.046	0.327	0.002	0.011	0.200	0.028	0.059	0.465	0.011	0.025	0.437	0.002	0.005	0.411
MAR	0.019	0.062	0.299	0.015	0.046	0.325	0.002	0.011	0.195	0.028	0.060	0.464	0.011	0.026	0.435	0.002	0.005	0.402
ILA	0.019	0.062	0.299	0.015	0.046	0.326	0.002	0.011	0.198	0.028	0.059	0.464	0.011	0.025	0.436	0.002	0.005	0.405

Note.—Since the values of dN/dS for 14 human populations from the 1000 Genomes Project are very similar, hence only representative populations are shown in the table. ASW, CEU, MXL, YRI, and CHB are, respectively, considered to be represented by African-Americans, Europeans, Americans, Africans, and Asians. The descriptions of the Population codes are provided in table 1.

region around the mutational site. This pattern of high homozygosity is not expected in an extended region if purifying selection operates. Therefore, extended haplotype homozygosity is a hallmark of positive selection.

Extended haplotype homozygosity (EHH) was not observed for *TLR2* or *TLR4* (figs. 2–4), although there is some variation in the patterns of haplotype homozygosity among Indian and other global populations. Such variability in patterns of EHH can be caused by variations in the sizes of the founding population and resultant genetic drift, especially if the founder sizes were small. Positive selection usually swamps the effects of other evolutionary forces and results in the appearance of

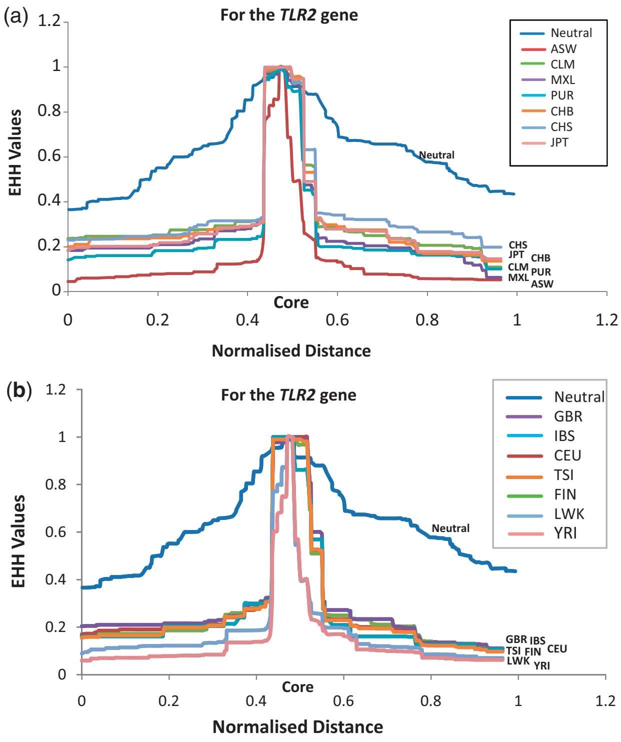

FIG. 2.—Extended haplotype homozygosity plots for the *TLR2* gene in (*a*) American and Asian populations and (*b*) African and European populations included in the 1000 Genomes Project.

EHH in a short period of time. As EHH was not observed for the two TLRs in most populations, we conclude that there are no compelling signs of positive selection operating on these genes; the possibility that negative or purifying selection has been operating is more likely.

Haplotype Networks Confirm the Presence of Excess of Rare Variants

Haplotype networks were constructed from the genotype data for each gene separately for the continental and Indian urban and tribal populations. These networks exhibit similar features; 1) very few high-frequency haplotype nodes; 2) large number of moderate and low frequency haplotype nodes connected to the high-frequency nodes resulting in local star-like phylogenies within each large network; and 3) the ancestral chimpanzee haplotype is ubiquitously found to be many mutational steps away from the human modal haplotypes (supplementary figs. S1–S4, Supplementary Material online). These features are, again, consistent with purifying selection. Further, the observation that the chimpanzee haplotypes are many mutational steps away from the human modal haplotypes may indicate that selection pressures unique to the humans may have operated since the separation of the human and chimpanzee lineages.

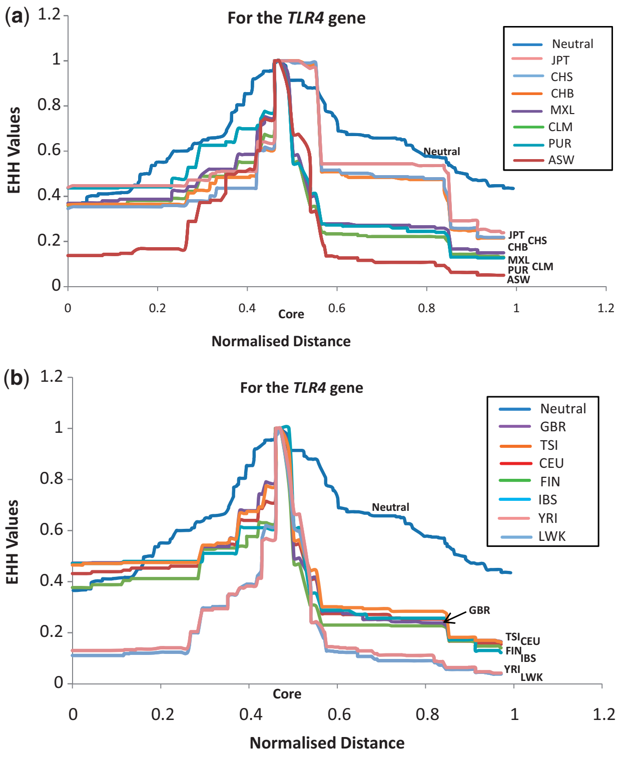

FIG. 3.—Extended haplotype homozygosity plots for the *TLR4* gene in (*a*) American and Asian populations and (*b*) African and European populations included in the 1000 Genomes Project.

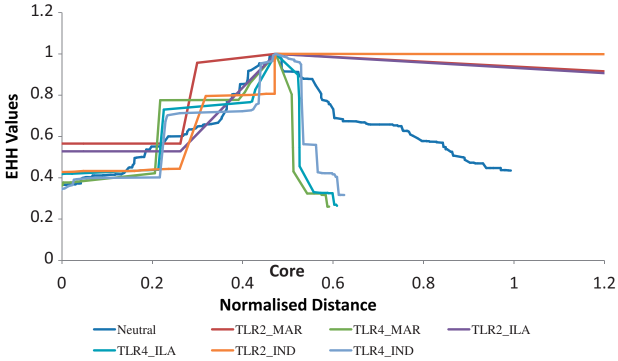

FIG. 4.—EHH plots of *TLR2* and *TLR4* in IND, MAR, and ILA populations compared with neutral expectations.

DNA Sequence Comparison with Orthologous Primate Species Confirms Conservation of Major Alleles Across Variant Loci

Human sequences from multiple populations were aligned with the chimpanzee, gibbon, and rhesus sequences for the *TLR2* and *TLR4* genes. The results provided evidence of DNA sequence conservation and showed that most (64% for *TLR2*

Table 5

DNA Sequence Conservation for Rare Mutations (MAF < 0.05) in Human Populations Compared with Some Ancestral Primate Reference Sequences for *TLR2* and *TLR4* Genes

Gene Name	No. of Loci[a] in 14 Human Populations	No. of Loci at Which a Minor Allele in Humans Is the Refseq Allele In			Total No. of Rare Alleles in Humans That Are Conserved	Total No. of Rare Alleles in Humans That Are New
		Chimp	Gibbon	Rhesus		
TLR2	115[b]	13 (11.3%)	12 (10.4%)	16 (13.9%)	41 (35.6%)	74 (64.4%)
TLR4	47[c]	1 (2.1%)	6 (12.8%)	3 (6.4%)	10 (21.3%)	37 (78.7%)

[a] Loci which have MAF < 0.05 and were not monomorphic in at least two populations out of 14 populations.
[b] DNA sequences adjacent to 22 human loci could not be aligned to Gibbon and 16 human loci could not be aligned to Rhesus reference sequences.
[c] DNA sequence adjacent to 1 human locus could not be aligned to Rhesus reference sequence.

and 79% for *TLR4*) of the rare alleles in these two genes found in human populations are new; the total number of rare variants being 115 for *TLR2* and 47 for *TLR4* with MAF less than 0.05 and not monomorphic in at least 2 out of 14 populations (table 5).

Discussion

In view of conflicting reports on the mode of selection operating on genes of the innate immune system, we have generated and systematically analyzed sequence variation data on two representative genes of the innate immune system, *TLR2* and *TLR4*. Both genes are members of the TLR family and are cell-surface receptors that recognize PAMPs of Gram-positive and Gram-negative pathogenic bacteria in humans. As India remains underrepresented in most global population genetic studies, such as HapMap and 1000 Genomes Project, we generated data on population groups representing distinct sociocultural and genetic diversity. We analyzed the Indian data in conjunction with DNA sequence data on these genes available on 14 populations included in the 1000 Genomes Project because we posited that previous, conflicting inferences on the mode of selection on the TLRs may have been due to analyses of only regional data. Analysis of these global data, we believe, has provided enhanced statistical power for inferring the mode of natural selection operating on the cell-surface TLRs of the innate immune system. These data were analyzed by using a set of robust statistical methods. Haplotype analysis has revealed that most populations harbor a large number of low-frequency haplotypes and few high-frequency modal haplotypes. Haplotype networks have star-like phylogenies centered around the modal haplotypes, which are all several mutational steps away from the chimpanzee reference haplotype, possibly reflecting new evolutionary pressures of natural selection subsequent to the separation of the chimpanzee lineage from the human lineage. Patterns of haplotype homozygosity do not reveal long and extended segments. The values for the neutrality test statistics are mostly negative. The ratios of nonsynonymous to synonymous changes are all less than unity. DNA sequence comparisons with multiple primate species have revealed near-conservation of alleles at variant loci in humans providing conclusive evidence that new mutations are quickly eliminated. These features are more consistent with purifying selection than other selection regimes.

The innate immune system in general and the cell-surface TLRs in particular, recognize motifs that are conserved across a broad range of pathogens. As a result, functional constraints are imposed on mutations. Newly arising mutations, therefore, are quickly removed from the population, providing genomic signatures that are consistent with purifying selection. By analyzing a large data set generated from diverse global populations using a consistent set of statistical methods, we have been able to provide substantial evidence in resolving the conflict regarding the selection regime that maintains sequence variation in genes of the innate immune system, particularly the cell-surface TLRs.

Supplementary Material

Supplementary tables S1–S9 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank all members of TCG-ISI Centre for Population Genomics, Kolkata, and the Centre for Genomic Application, New Delhi, India, for logistical and infrastructural support. They are also thankful to Sujata Dhar for computational assistance. They express their sincere gratitude to the members of the 1000 Genomes Project for making their data publicly accessible through their database. This work was partially supported by the Ministry of Science & Technology, Government of India, grant under the India-Spain S&T Co-operation.

Literature Cited

Areal H, Abrantes J, Esteves PJ. 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol Biol. 11:368.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Barreiro LB, et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5: e1000562.

Budulac SE, et al. 2012. Toll-like receptor (TLR2 and TLR4) polymorphisms and chronic obstructive pulmonary disease. PLoS One 7:e43124.

Daszak P, Cunningham AA, Hyatt AD. 2000. Emerging infectious diseases of wildlife—threats to biodiversity and human health. Science 287: 443–449.

Diamond J. 1997. Guns, germs, and steel: the fates of human societies. New York: Norton.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. Nature 418:700–707.

Elson G, Dunn-Siegrist I, Daubeuf B, Pugin J. 2007. Contribution of Toll-like receptors to the innate immune response to Gram-negative and Gram-positive bacteria. Blood 109:1574–1583.

Epstein PR. 1995. Emerging diseases and ecosystem instability: new threats to public health. Am J Public Health. 85:168–172.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 10:564–567.

Ferrer-Admetlla A, et al. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol. 181: 1315–1322.

Fiennes R. 1978. Zoonoses and the origins and ecology of human disease. London: Academic Press.

Haldane JBS. 1949. Disease and evolution. Ric Sci Suppl A. 19:68–76.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet. 10(9):639–650.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Ito M, et al. 2012. The association of Toll-like receptor 4 gene polymorphisms with the development of emphysema in Japanese subjects: a case control study. BMC Res Notes. 5:36.

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

Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Mcneill WH. 1989. Plagues and peoples, 2nd. New York: Anchor Books.

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 16: 1215.

Mukherjee S, Sarkar-Roy N, Wagener DK, Majumder PP. 2009. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc Natl Acad Sci U S A. 106:7073–7078.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nischalke HD, et al. 2012. The Toll-like receptor 2 (TLR2) -196 to -174 del/ins polymorphism affects viral loads and susceptibility to hepatocellular carcinoma in chronic hepatitis C. Int J Cancer. 130:1470–1475.

Parham P. 2003. The unsung heroes. Nature 423:20.

Purcell S, et al. 2007. PLINK: a tool set for whole genome association and population-based linkage analyses. Am J Hum Genet. 81:559–575.

Rosenberg NA, et al. 2002. Genetic structure of human populations. Science 298:2381–2385.

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

Sabeti PC, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sato K, et al. 2012. A single nucleotide polymorphism in 3′-untranslated region contributes to the regulation of Toll-like receptor 4 translation. J Biol Chem. 287:25163–25172.

Seabury CM, et al. 2010. Diversity and evolution of 11 innate immune genes in *Bos taurus taurus* and *Bos taurus indicus* cattle. Proc Natl Acad Sci U S A. 107:151–156.

Schrag SJ, Wiener P. 1995. Emerging infectious disease: what are the relative roles of ecology and evolution? Trends Ecol Evol. 10:319–324.

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

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68: 978–989.

Takeuchi O, et al. 1999. Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. Immunity 11:443–451.

Texereau J, et al. 2005. The importance of Toll like receptor 2 polymorphisms in severe infections. Clin Infect Dis. 41:S408–S415.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Wlasiuk G, Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. Mol Biol Evol. 27:2172–2186.

Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. Nature 447:279–283.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Associate editor: Bill Martin