Current Biology, Vol. 15, 1022–1027, June 7, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.04.050

Pathogen-Driven Selection and Worldwide HLA Class I Diversity

Franck Prugnolle,[1],* Andrea Manica,[2]
Marie Charpentier,[3] Jean François Guégan,[4]
Vanina Guernier,[4,5] and François Balloux[1]
[1] Theoretical and Molecular Population Genetics Group
Department of Genetics
University of Cambridge
Downing Street
Cambridge CB2 3EH
United Kingdom
[2] Evolutionary Ecology Group
Department of Zoology
University of Cambridge
Downing Street
Cambridge CB2 3EJ
United Kingdom
[3] Centre d'Ecologie Fonctionnelle et Evolutive-Centre National de la Recherche Scientifique (CNRS)
1919 route de Mende
34293 Montpellier cedex 05
France
[4] Génétique et Evolution des Maladies Infectieuses
Unité 2724 Institut de Recherche et Développement (IRD)-CNRS
Equipe Evolution des Systèmes Symbiotiques
911 avenue Agropolis
BP 64501
34394 Montpellier cedex 05
France
[5] Unité ESPACE S140
Institut de Recherche et Développement
Maison de la Télédétection
500 rue J.F. Breton
34093 Montpellier cedex 05
France

Summary

The human leukocyte antigen (HLA; known as MHC in other vertebrates) plays a central role in the recognition and presentation of antigens to the immune system and represents the most polymorphic gene cluster in the human genome [1]. Pathogen-driven balancing selection (PDBS) has been previously hypothesized to explain the remarkable polymorphism in the HLA complex, but there is, as yet, no direct support for this hypothesis [2, 3]. A straightforward prediction coming out of the PDBS hypothesis is that populations from areas with high pathogen diversity should have increased HLA diversity in relation to their average genomic diversity. We tested this prediction by using HLA class I genetic diversity from 61 human populations. Our results show that human colonization history explains a substantial proportion of HLA genetic diversity worldwide. However, between-population variation at the HLA class I genes is also positively correlated with local pathogen richness (notably for the HLA B gene), thus providing support for the PDBS hypothesis. The proportion of variations explained by pathogen richness is higher for the HLA B gene than for the HLA A and HLA C genes. This is in good agreement with both previous immunological and genetic data suggesting that HLA B could be under a higher selective pressure from pathogens.

Results and Discussion

There are several lines of evidence suggesting that genetic diversity at the HLA (MHC) complex is driven and maintained by a process of diversifying selection [4–10]: (1) A large number of alleles have been observed in nearly all vertebrates studied to date [11], even in species with virtually no genetic diversity at other loci [8]; (2) alleles are distributed more or less evenly within populations [7]; and (3) the rate of nonsynonymous substitutions in the antigen binding region of the molecule is higher than that of synonymous substitutions [4, 5].

The mechanisms that have been proposed to explain the evolution of HLA (MHC) polymorphism in natural populations vary from MHC-dependent mate selection and preferential abortion to various pathogen-driven selection pressures (see [2] for an extensive review). Pathogen-driven selection is expected to operate when specific alleles are favored because of their ability to provide protection from different pathogen species or strains. There is much evidence, both from immunology and genetics, that points toward pathogen-driven selection for HLA diversity. For example, several studies report that specific HLA (MHC) alleles provide increased resistance against single pathogens [12–17]. Moreover, more diverse individuals at this gene complex tend to have higher fitness when challenged with a variety of pathogens, suggesting the presence of balancing selection (i.e., heterozygotes have higher fitness than either homozygote genotype) [18–23]. However, although all these elements are compatible with PDBS [2, 18, 24], none of these findings is really sufficient to conclude that extreme polymorphism of HLA (MHC) genes in natural populations is driven by the diversity of pathogens encountered [3, 16, 25, 26].

A testable prediction from the PDBS hypothesis is that human populations from areas with high pathogen diversity should have increased HLA diversity in relation to their neutral diversity. This is because individuals with higher variability at the HLA genes are predicted to cope with a higher number of pathogens. In this paper, we test this prediction with three different sources of information: (1) the genetic diversity at HLA class I genes (A, B, and C) from 61 human populations distributed over the world (see the Supplemental Data available with this article online for more details on methods and populations); (2) estimates of intracellular pathogen richness (the total number of intracellular human disease agents known from each country where HLA-

*Correspondence: prugnolle@yahoo.fr

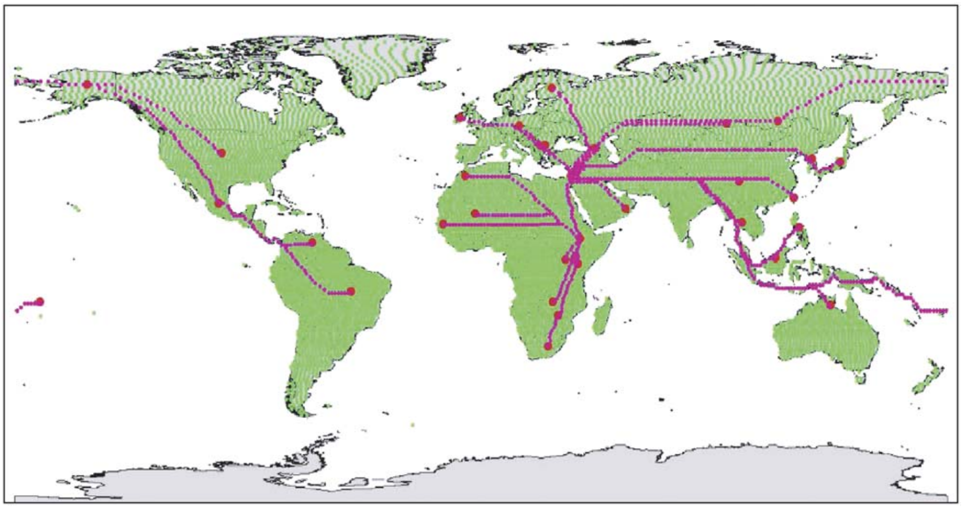

Figure 1. Shortest Routes (in Purple) through Landmasses and Specified Land Bridges between the 61 Populations Analyzed (Red Dots) and a Hypothetical East African Origin

Geographic distances have been computed as paths connecting vertices (in green) on land with an algorithm based on graph theory (see [28]).

typed populations originate (see Supplemental Data); and (3) the geographic distance of each population from East Africa along likely ancient colonization routes (see Supplemental Data; Figure 1), to control for the effect of past colonization history.

The diversity of a gene under selection in any population is not only a function of selective factors acting specifically on that locus, but also of demography that similarly affects all loci in the genome [27]. We have recently shown that the genetic diversity observed in current human populations at neutral markers ($H_S{}^n$) is extraordinarily well correlated with the geographic distance through landmasses between those populations and East Africa (r^2 = 85%) [28]. Here, we further improve on this correlation (r^2 = 87%) by applying the following transformation to neutral genetic diversity: $H_S{}^{n*} = \mathrm{Log}\,[H_S{}^n/(1-H_S{}^n)]$. Such a strong relationship allows us to use geographic coordinates as proxies for the neutral genetic diversity of populations situated at any point on the globe. Such information can be further exploited to disentangle the effect of past colonization history and potential natural selection that may have shaped the current apportionment of genetic diversity at any locus in the genome.

Worldwide variation in HLA diversity ($H_S{}^{HLA*}$) is significantly correlated to geographic distance from East Africa (the distance being computed through landmasses), with populations farther from Ethiopia being characterized by lower genetic variability (Table 1; Figure 2). Geographic distance and, hence, human colonization history explains only between 17% to 39% of the genetic diversity observed at HLA genes (Table 1; Figure 2). Although sampling effects could be invoked to explain why the correlation observed for HLA genes is lower than the one previously obtained with 377 neutral microsatellite markers [28], selective pressures imposed by pathogens also seem important. Indeed, the remaining proportion of HLA diversity, which is not explained by the human colonization history, is significantly correlated with pathogen richness (Table 1). Human populations that are exposed to a more diverse array of pathogens show higher HLA diversity than those exposed to fewer pathogens (Table 1). The corre-

Table 1. Regression Analyses for the Relationship among HLA Genetic Diversity ($H_S{}^{HLA*}$), Geographic Distance from Africa (*Dist. Africa*), and Intracellular Pathogen Species Richness

		HLA A	HLA B	HLA C
n		61	59	48
Model I				
Dist. Africa	r^2	39%***	17%***	35%***
Path. Rich.	r^2	5%*	10.5%**	2.6% ns
Model II				
Dist. Africa	r^2	39%***	17%***	35%***
Viruses	r^2	8%**	11%**	4.7% ns
Model III				
Dist. Africa	r^2	39%***	17%***	35%***
Bacteria O	r^2	1.5% ns	4% ns	2% ns
Model IV				
Dist. Africa	r^2	39%***	17%***	35%***
Bacteria F	r^2	0.2% ns	6%*	0% ns
Model V				
Dist. Africa	r^2	39%***	17%***	35%***
Protozoa	r^2	4%*	6%*	0.5% ns

Regression models I–V were fitted independently according to the procedure detailed in Statistical Analyses (see Supplemental Data). Path. Rich, Bacteria O, and Bacteria F denote species richness of all intracellular pathogens and obligate and facultative intracellular bacteria, respectively. n represents the number of populations genotyped, and r^2 the proportion of variance explained by each independent variable. All the slopes of regressions between pathogen richness and HLA class I diversity were positive. P values for *F* tests: *** < 0.001; ** < 0.01; * < 0.05; ns, non-significant.

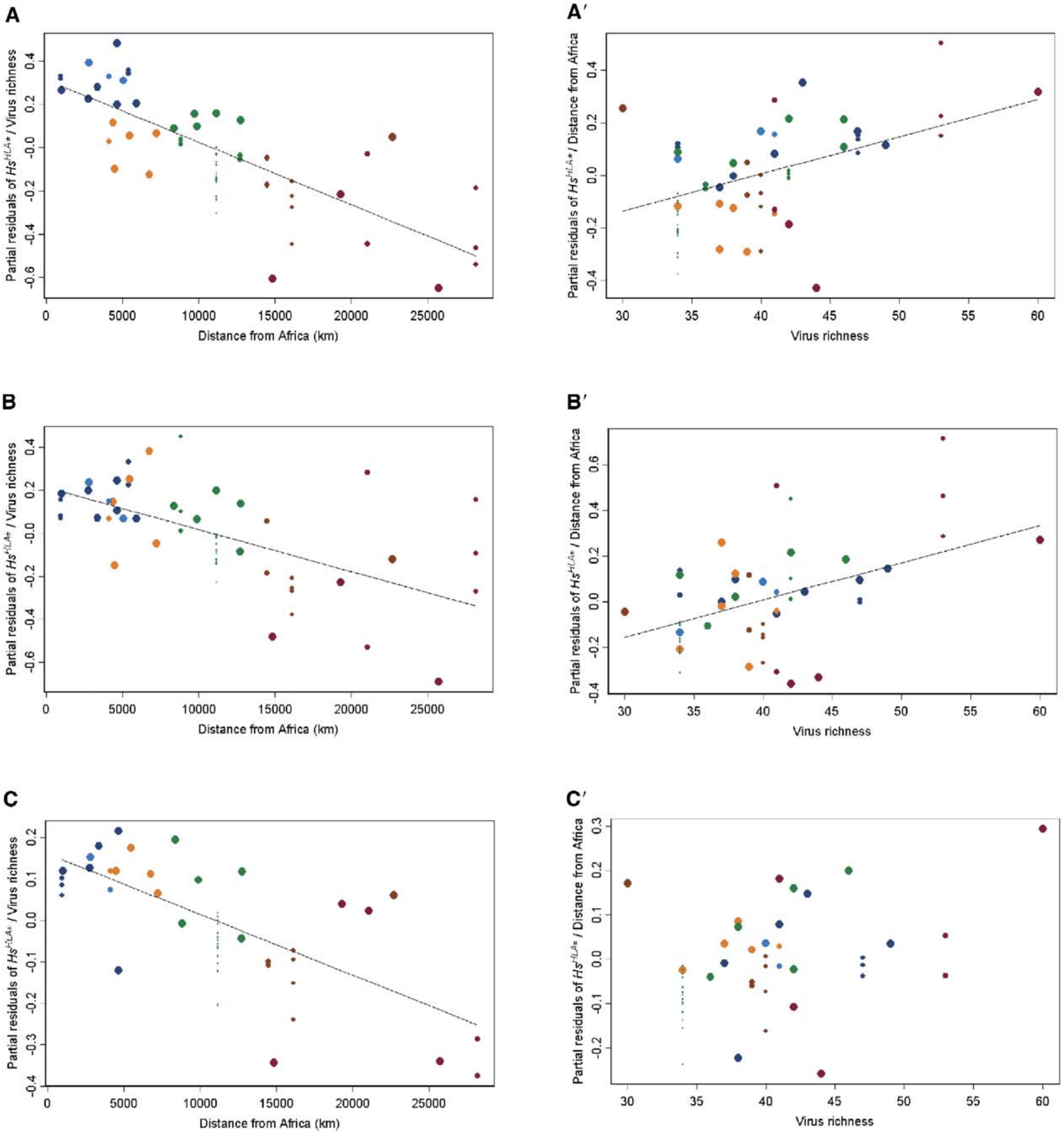

Figure 2. Partial Residuals of H_S^{HLA*} against Distance from Africa and Virus Richness, the Two Predictors from Our Model

(A–C) Partial residuals of H_S^{HLA*} (after fitting Virus Richness) versus the Distance from Africa (in km) for HLA A (A), HLA B (B), and HLA C (C) genes.

(A′–C′) Partial residuals of H_S^{HLA*} (after fitting Distance from Africa) versus Virus Richness for HLA A (A′), HLA B (B′), and HLA C (C′) genes. More details about these regressions are given in Table 1 (Model II) for all HLA class I genes. Lines represent partial linear regressions. The size of the points is proportional to the weights applied to H_S^{HLA*} in the model. The different colors correspond to the different ethnic groups (dark blue, African populations; sky blue, Middle East populations; orange, European populations; green, East Asian populations; brown, Oceanian populations; and purple, American populations).

lation is significant for HLA A and B genes. This relationship is mainly driven by virus richness (Table 1), which explains a higher proportion of the remaining variations for HLA B (11%) than for HLA A (8%) and HLA C (4.7%). Several lines of evidence reveal these results to be very robust. First, we obtain similar proportions when we reanalyze the HLA A and B datasets after subsetting them to include only the populations that are available for HLA C (n = 48; see Table S2). Second, alternative geographic models (including latitude as another potential explanatory variable) do not change the results qualitatively either (Table 2). Finally,

Table 2. Regressions for the Internal Controls and the Alternative Geographic Model (see Statistical Analyses)

Internal Controls			
HLA DRB1	Dist. Africa	r^2	47% (–)***
	Viruses	r^2	4% (–) ns
Locus ms D12S1638	Dist. Africa	r^2	57% (–)***
	Viruses	r^2	0.2% (–) ns
Locus ms D20S103	Dist. Africa	r^2	31% (–)***
	Viruses	r^2	<0.1% (–) ns
Locus ms D10S1412	Dist. Africa	r^2	67% (–)***
	Viruses	r^2	1.3% (–) ns
Locus ms D16S3396	Dist. Africa	r^2	22% (–)***
	Viruses	r^2	0.3% (+) ns
Locus ms D4S3243	Dist. Africa	r^2	14% (–)**
	Viruses	r^2	2.8% (–) ns
Locus ms NA-D6S-1	Dist. Africa	r^2	16% (–)**
	Viruses	r^2	1.6% (–) ns
Locus ms D6S1009	Dist. Africa	r^2	16% (–)**
	Viruses	r^2	0.6% (–) ns
Locus ms D5S2501	Dist. Africa	r^2	40% (–)***
	Viruses	r^2	5% (+) ns
Locus ms D10S1430	Dist. Africa	r^2	17% (–)**
	Viruses	r^2	0.2% (+) ns
Locus ms D21S1437	Dist. Africa	r^2	16% (–)**
	Viruses	r^2	0.7% (+) ns
Alternative Geographic Model			
HLA A	Dist. Africa	r^2	39% (–)***
	Abs. Lat.	r^2	5% (–)*
	Viruses	r^2	7% (–)**
HLA B	Dist. Africa	r^2	17% (–)***
	Abs. Lat.	r^2	<0.1% (+) ns
	Viruses	r^2	11% (+)**
HLA C	Dist. Africa	r^2	35% (–)***
	Abs. Lat.	r^2	<0.1% (+) ns
	Viruses	r^2	4.7% (–) ns

Internal Controls: Regressions between genetic diversity at one HLA locus of class II (HLA DRB1) and at ten microsatellite markers (ms) randomly chosen from the 377 used in Rosenberg et al. [35], geographic distance from Africa (Dist. Africa), and virus species richness. Alternative Geographic Model: The absolute value of the latitude (Abs. Lat.) of the countries from which populations come is entered as another potential explanatory variable of the HLA class I genetic diversity. r^2 represents the proportion of variance explained by each independent variable. (–) or (+) indicates the sign of the regression slope. P values for *F* tests: *** < 0.001; ** < 0.01; * < 0.05; ns, non-significant.

as we expected, when the same tests are performed for genes that are not expected to be under the selection of intracellular pathogens (one locus of HLA class II and ten randomly selected microsatellite markers), no significant relationship was ever found with virus richness (Table 2).

Our results therefore support the PDBS hypothesis and further suggest that pathogens (notably viruses) might exert a stronger selection pressure on the HLA B gene than on HLA A and HLA C genes. These conclusions are supported by previous studies. First, it has been demonstrated that the strongest balancing selection operates at the HLA B locus (selection coefficient s = 4.2%; [29]) followed by the HLA A (s =1.5%; [29]) and HLA C loci (s = 0.26%; [29]; see also [30]), which is in good agreement with the proportion of variance explained by pathogen richness for the different HLA class I genes in the present study (Table 1). Second, it has been shown that, despite apparently similar roles in pathogen defense (notably against viruses), there are well-established differences between HLA class I loci [10]. HLA C alleles are expressed at lower levels on the cell surface than HLA A and HLA B [31], and, thus, it is not unexpected that HLA C diversity is the least driven by pathogen selection of the three HLA class I loci. For the difference in selection between HLA A and B, it has been very recently demonstrated that the HLA B gene could play a larger role in the successful containment of viral infection (notably HIV infection) and would therefore be under higher diversifying selection pressure from the intracellular disease agents (notably viruses) than the HLA A gene [32]. Finally, many studies report that some specific HLA (MHC) alleles provide increased resistance against single pathogens [12–17] or that heterozygous individuals are better at coping with particular pathogens, especially when they are challenged by a mixture of infectious agents [18–23].

It is interesting to note that the correlation between HLA diversity (corrected for colonization history) and pathogen richness is detected despite the fact that we used a very conservative measurement for the selective pressure induced by pathogen communities (i.e., simply counting the number of disease agents recorded in each country irrespective of their burden). Not accounting for prevalence translates into some diseases being present in essentially all populations and thus not having any statistical weight. This approach further underestimates the true effect of many diseases, such as influenza or malaria, comprising heterogeneous mixes of pathogen strains that may have different HLA associations.

Obviously, an implicit assumption behind our interpretation of these results as supportive for the PDBS hypothesis is that both current patterns of human genetic diversity and present distribution of pathogens on Earth reflect the conditions that may have shaped the evolutionary relation between pathogen richness and HLA polymorphism in the past. In particular, we assume (1) that current pathogen richness provides a reliable picture of the selective landscape experienced by human populations in the past and (2) that the current distribution of human genetic diversity has not been affected very much by the recent increase in human migration.

Numerous countries have recently undergone what is known as the epidemiological transition (some 300 years ago in some developed countries and less than 80 years ago for underdeveloped countries) [33]. This transition corresponds to major changes whereby parasitic-disease mortality decreased, reducing the selective pressure imposed on human populations by infectious diseases [33]. However, despite medical progress, global selective pressure imposed by pathogens is still very high because infectious diseases continue to be a major cause of mortality—they are responsible for 48% of deaths worldwide in people under the age of 45 [34]. Although the relative fitness costs owing to individual pathogen species have greatly evolved over the last centuries, with some diseases previously causing extreme mortality now under control and others expanding their range (emerging infectious diseases such as HIV), it is unlikely that the relative number of pathogens per country, irrespective of their prevalence, has greatly changed over recent times. Equally, recent population admixture linked to an increase of migration is unlikely to have had major effects on a worldwide scale (at least for native populations such as those considered in the present study), as shown by the excellent relationship between geographic distance to Ethiopia and neutral genetic diversity [28].

In conclusion, our results add strong support to the PDBS hypothesis. To date, many studies on HLA have focused on variation within single populations [16]. Although such an approach has proved very successful at detecting evidence for natural selection, it is not well suited for inferring the nature of the selective agent. In this paper, we approached the question in a comparative manner by contrasting patterns of variation at HLA genes, the effect of historical migrations, and selection pressures across populations. This approach enabled us to show that while human colonization history has been important in shaping the present patterns observed at HLA genes, the diversity of pathogens has also been important in driving and maintaining the genetic diversity at HLA genes. We believe that this study adds an exciting example to the few cases of natural selection, acting on the human genome, for which the underlying selective factors are identified.

Supplemental Data

Supplemental Data including a full description of the databases and methods used and some supplemental results and analyses are available at http://www.current-biology.com/cgi/content/full/15/11/1022/DC1/.

Acknowledgments

We thank Matt Bell, Bruno Bucheton, Tim Coulson, Thierry De Meeûs, David Hosken, François Renaud, Frederic Thomas, Elizabeth Tyler, and four referees for comments on the manuscript. F.P. is supported by a Lavoisier Fellowship from the Ministère Français des Affaires Etrangères. J.F.G. is supported by IRD and CNRS. V.G. is supported by the Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT). F.B. acknowledges support from the Biotechnology and Biosciences Research Council (BBSRC).

Received: February 4, 2005
Revised: April 20, 2005
Accepted: April 21, 2005
Published: June 7, 2005

References

1. Hughes, A.L., and Yeager, M. (1998). Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. *32*, 415–435.
2. Apanius, V., Penn, D., Slev, P.R., Ruff, L.R., and Potts, W.K. (1997). The nature of selection on the major histocompatibility complex. Crit. Rev. Immunol. *17*, 179–224.
3. Jeffery, K.J.M., and Bangham, C.R.M. (2000). Do infectious diseases drive MHC diversity? Microbes Infect. *2*, 1335–1341.
4. Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. Nature *335*, 167–170.
5. Hughes, A.L., and Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class-II loci—evidence for overdominant selection. Proc. Natl. Acad. Sci. USA *86*, 958–962.
6. Hedrick, P.W., Parker, K.M., Miller, E.L., and Miller, P.S. (1999). Major histocompatibility complex variation in the endangered Przewalski's horse. Genetics *152*, 1701–1710.
7. Garrigan, D., and Hedrick, P.W. (2003). Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. Evolution Int. J. Org. Evolution *57*, 1707–1722.
8. Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., and Wayne, R.K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc. Natl. Acad. Sci. USA *101*, 3490–3494.
9. Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., et al. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens *63*, 293–325.
10. Parham, P., and Ohta, T. (1996). Population biology of antigen presentation by MHC class I molecules. Science *272*, 67–74.
11. Bernatchez, L., and Landry, C. (2003). MHC studies in non-model vertebrates: What have we learned about natural selection in 15 years? J. Evol. Biol. *16*, 363–377.
12. Hill, A.V., Allsopp, C.E.M., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J., and Greenwood, B.M. (1991). Common West African HLA antigens are associated with protection from severe malaria. Nature *352*, 595–600.
13. Thursz, M.R., Kwiatkowski, D., Allsopp, C.E.M., Greenwood, B.M., Thomas, H.C., and Hill, A.V.S. (1995). Association between an MHC class-II allele and clearance of hepatitis-B virus in the Gambia. N. Engl. J. Med. *332*, 1065–1069.
14. Paterson, S., Wilson, K., and Pemberton, J.M. (1998). Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population (Ovis aries L.). Proc. Natl. Acad. Sci. USA *95*, 3714–3719.
15. Godot, V., Harraga, S., Beurton, I., Tiberghien, P., Sarciron, E., Gottstein, B., and Vuitton, D.A. (2000). Resistance/susceptibility to *Echinococcus multilocularis* infection and cytokine profile in humans. II. Influence of the HLA B8, DR3, DQ2 haplotype. Clin. Exp. Immunol. *121*, 491–498.
16. Meyer, D., and Thomson, G. (2001). How selection shapes variation of the human major histocompatibility complex: A review. Ann. Hum. Genet. *65*, 1–26.
17. Trachtenberg, E., Korber, B., Sollars, C., Kepler, T.B., Hraber, P.T., Hayes, E., Funkhouser, R., Fugate, M., Theiler, J., Hsu, Y.S., et al. (2003). Advantage of rare HLA supertype in HIV disease progression. Nat. Med. *9*, 928–935.
18. Doherty, P.C., and Zinkernagel, R.M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature *256*, 50–52.
19. McClelland, E.E., Penn, D., Fujinami, R.S., and Potts, W.K. (2000). MHC heterozygote advantage under coinfection with Salmonella and Theiler's virus. FASEB J. *14*, A1054–A1054.
20. Penn, D.J., Damjanovich, K., and Potts, W.K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc. Natl. Acad. Sci. USA *99*, 11260–11264.
21. McClelland, E.E., Penn, D.J., and Potts, W.K. (2003). Major histocompatibility complex heterozygote superiority during coinfection. Infect. Immun. *71*, 2079–2086.
22. Thursz, M.R., Thomas, H.C., Greenwood, B.M., and Hill, A.V.S. (1997). Heterozygote advantage for HLA class-II type in hepatitis B virus infection. Nat. Genet. *17*, 11–12.
23. Carrington, M., Nelson, G.W., Martin, M.P., Kissner, T., Vlahov, D., Goedert, J.J., Kaslow, R., Buchbinder, S., Hoots, K., and O'Brien, S.J. (1999). HLA and HIV-1: Heterozygote advantage and B*35-Cw*04 disadvantage. Science *283*, 1748–1752.
24. Clarke, B.C., and Kirby, D.R.S. (1966). Maintenance of histocompatibility polymorphism. Nature *211*, 999–1000.
25. Klein, J., and Ohuigin, C. (1994). Mhc polymorphism and parasites. Philos. Trans. R. Soc. Lond. B Biol. Sci. *346*, 351–357.
26. Hedrick, P.W., and Kim, T.J. (2000). Genetics of complex polymorphisms: Parasites and maintenance of the major histocompatibility complex variation. In Evolutionary Genetics: From Molecules to Morphology, R.S. Singh and C.B. Krimbas, eds. (Cambridge: Cambridge University Press), pp. 204–234.
27. Bamshad, M., and Wooding, S.P. (2003). Signatures of natural selection in the human genome. Nat. Rev. Genet. *4*, 99–111.

28. Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Curr. Biol. *15*, R159–R160.
29. Satta, Y., Ohuigin, C., Takahata, N., and Klein, J. (1994). Intensity of natural selection at the major histoincompatibility complex loci. Proc. Natl. Acad. Sci. USA *91*, 7184–7188.
30. Slatkin, M., and Muirhead, C.A. (2000). A method for estimating the intensity of overdominant selection from the distribution of allele frequencies. Genetics *156*, 2119–2126.
31. Snary, D., Barnstable, C.J., Bodmer, W.F., and Crumpton, M.J. (1977). Molecular structure of human histocompatibility antigens: The HLA-C series. Eur. J. Immunol. *7*, 580–585.
32. Kiepiela, P., Leslie, A.J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K.J., Hilton, L., et al. (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature *432*, 769–774.
33. Omran, A.R. (1982). Epidemiologic transition. In International Encyclopedia of Populations, J.A. Ross, ed. (London: The Free Press), pp. 172–183.
34. Kapp, C. (1999). WHO warns of microbial threat. Lancet *353*, 2222.
35. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.