ARTICLES

nature medicine

Advantage of rare HLA supertype in HIV disease progression

Elizabeth Trachtenberg[1,7], Bette Korber[2,3,7], Cristina Sollars[1], Thomas B Kepler[3,4], Peter T Hraber[3], Elizabeth Hayes[1], Robert Funkhouser[2,3], Michael Fugate[2], James Theiler[2], Yen S Hsu[1], Kevin Kunstman[5], Samuel Wu[5], John Phair[5], Henry Erlich[1,6] & Steven Wolinsky[5]

The highly polymorphic human leukocyte antigen (HLA) class I molecules help to determine the specificity and repertoire of the immune response. The great diversity of these antigen-binding molecules confers differential advantages in responding to pathogens, but presents a major obstacle to distinguishing *HLA* allele–specific effects. HLA class I supertypes provide a functional classification for the many different *HLA* alleles that overlap in their peptide-binding specificities. We analyzed the association of these discrete HLA supertypes with HIV disease progression rates in a population of HIV-infected men. We found that HLA supertypes alone and in combination conferred a strong differential advantage in responding to HIV infection, independent of the contribution of single *HLA* alleles that associate with progression of the disease. The correlation of the frequency of the HLA supertypes with viral load suggests that HIV adapts to the most frequent alleles in the population, providing a selective advantage for those individuals who express rare alleles.

The *HLA* loci encode two distinct classes of highly polymorphic cell surface glycoproteins that bind and present processed antigenic peptides to T cells of the immune system[1]. HLA class I molecules present endogenous antigen, synthesized and processed in the infected cell cytoplasm by intracellular bacteria and viruses, to $CD8^+$ cytotoxic T lymphocytes (CTLs) that then kill the infected cell[2]. HLA class II molecules display exogenously derived epitopes on the surface of antigen-presenting cells for immune recognition by $CD4^+$ helper T cells, which then coordinate the immune response against an invading pathogen[3]. Polymorphisms that cluster within the antigenic-peptide binding cleft provide a range of divergent peptide-binding specificities that determine the spectrum of epitopes that are bound and presented by a particular HLA molecule.

The extensive polymorphism at the *HLA* loci is thought to have arisen through natural selection by infectious diseases, operating on the diversity generated by mutation, gene conversion and recombination[4,5]. At a population level, genetic diversity of the *HLA* loci is maintained by overdominant selection relating to enhanced antigenic peptide-binding capacity, and therefore resistance to infectious disease[6,7]. Individuals heterozygous at *HLA* loci are capable of presenting a broader array of pathogen-derived peptides, resulting in a more diverse CTL repertoire and the ability to resist a greater breadth of infectious pathogens. Thus, the great diversity of *HLA* alleles in a population ensures that no single pathogen can decimate the entire population. Nevertheless, an epidemic infectious disease such as AIDS can place populations under strong selection pressure from a single pathogen[8,9]. Such evolutionary pressure tends to increase the frequency of any *HLA* allele that provides better immunity against the pathogen, and thereby influences infectious disease susceptibility and mortality[4,10–14]. In turn, microbes adapt to the host by mutating the epitopes targeted by the HLA-directed immune response. For epidemic infectious diseases, frequency-dependent selection could provide a selective advantage for those individuals who express rare alleles, as pathogens are more likely to develop mechanisms to evade the immune response mediated by common *HLA* genotypes[8,15]. As these rare alleles gradually increase in frequency in the population through increased mortality, they themselves eventually become prime targets for microbial adaptation[15,16].

The incidence and clinical outcome of HIV infection are influenced by differences in viral strains and host genetic factors[17]. HLA associations with HIV disease have been somewhat inconsistent[18]. Survival analyses suggest that overall heterozygosity at *HLA* class I loci confers relative resistance to progression from HIV infection to AIDS[19,20]. Other reports show associations of particular *HLA* genotypes with HIV disease progression and transmission rates[18], but these studies can be complicated by multiple test issues, linkage disequilibrium, cohort effects and different measures of outcome. The extreme polymorphism of the six *HLA* class I and class II genes[5,21] complicates the attribution of specific alleles with the outcome of disease, such that collecting samples of the size needed for definitive results is often not feasible. Combinations of beneficial and

[1]Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way Oakland, California 94609, USA. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [3]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [4]Duke University Medical Center, Durham, North Carolina 27708, USA. [5]The Feinberg School of Medicine, Northwestern University, 676 North Saint Clair Street, Chicago, Illinois 60611, USA. [6]Roche Molecular Systems, 1145 Atlantic Avenue, Alameda, California 94501, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to S.W. (s-wolinsky@northwestern.edu).

detrimental *HLA* alleles can have opposing and counterbalancing effects in the host; thus, the involvement of multiple loci, the diploid genome and linkage disequilibrium can further complicate the analysis.

Here we conducted a population-based association study of men enrolled in the Chicago component of the Multicenter AIDS Cohort Study (MACS) to estimate the influence of *HLA* loci on progression of HIV disease. We capitalized on the significant overlap among the peptide binding motifs recognized by distinct HLA class I molecules, allowing functional classification of HLA supertypes based on the preferred F- and B-pocket contact residues of the antigenic peptides to which they bind[22]. The biological relevance of this classification scheme is supported by a growing body of evidence for cross-presentation of peptide-binding motifs by different HLA molecules assigned to a discrete supertype[10–14]. Because each supertype is relatively common in all major ethnic groups, the nine supertypes together cover most of the *HLA-A* and *HLA-B* alleles found in populations[22]. We addressed the statistical issues resulting from multiple tests that are inherent to studies of associations of HLA with disease using several new strategies.

Table 1 ***HLA*** **alleles previously documented to be associated with rates of progression to AIDS**

Specific HLA alleles associated with rapid CD4+ T-cell decline or high viral load				
	CD4+ T-cell slope *P* value	Viral load *P* value	Median (IR)	Associated outcome
*A*24*	0.0143	0.3904	–0.24 (–0.35 to –0.13)	CD4 slope
No *A*24*			–0.18 (–0.28 to –0.09)	CD4 slope
*B*37*	0.0017	0.2337	–0.32 (–0.47 to –0.28)	CD4 slope
No *B*37*			–0.18 (–0.29 to –0.09)	CD4 slope
*B*56*	0.9238	0.0092	91,180 (7,340 to 123,600)	Viral load
No *B*56*			15,280 (4,735 to 43,480)	Viral load
Specific HLA alleles associated with slow CD4+ T-cell decline or low viral load				
	CD4+ T-cell slope *P* value	Viral load *P* value	Median (IR)	Associated outcome
*B*57*	0.1615	<0.00001	3,411 (716 to 12,860)	Viral load
No *B*57*			17,330 (5,990 to 51,130)	Viral load
*A*32*	0.0238	0.1392	–0.12 (–0.23 to –0.08)	CD4 slope
No *A*32*			–0.19 (–0.30 to –0.10)	CD4 slope
*C*08*	0.0594	0.0065	7,898 (2,671 to 25,315)	Viral load
No *C*08*			12,510 (5,035 to 51,329)	Viral load
*DPB1*01*	0.0455	0.6548	–0.13 (–0.28 to –0.06)	CD4 slope
No *DPB1*01*			–0.19 (–0.29 to –0.11)	CD4 slope

HLA alleles associated with rapid CD4+ T-cell decline or high viral load, or slow CD4+ T-cell decline or low viral load. IR, interquantile range.

Our data confirm some of the previously reported *HLA* allelic associations, and moreover show that HLA supertypes, singly and in combination, confer a differential advantage in responding to HIV infection. Consistent with the rare-allele advantage model of frequency-dependent selection, we show that the frequency of a particular supertype in this HLA-diverse population correlates with the level of plasma viral RNA at the set point (viral load). These results support a fundamental role for both overdominance and frequency-dependent selection relating to the generation and maintenance of polymorphism at the *HLA* loci and validate the differential advantage they confer in responding to infectious disease.

RESULTS

HLA allelic associations with disease progression

We first compared our data with previously identified associations of *HLA* alleles[18] and HIV-related disease progression rates, being mindful of combinations of alleles that could confound the analysis. Our four-digit typing was reduced to a two-digit HLA typing for a more direct comparison with the previous literature. The *P* values are based on a nonparametric rank sum test comparing distributions of values for people who carry an *HLA* allele with those that do not. Given that we were corroborating prior observations in the literature, these values are not corrected for multiple tests. The results shown in Table 1 support several previous reports of the association of *HLA* alleles with the rate of progression of the disease[18,19,23]. *HLA-A*24* and *HLA-B*37* were associated with rapid CD4+ T-cell decline and *HLA-B*56* with high viral loads (**Table 1**). Conversely, *HLA-B*57* and *HLA-C*08* were associated with low viral loads, and *HLA-A*32* and *HLA-DPB1*01* with slow CD4+ T-cell decline (**Table 1**). In no case did we confirm a statistically significant association for both viral load and rate of CD4+ T-cell decline with *HLA* ($P < 0.05$), but this may simply result from analysis of inherently noisy data with limited sample size, or it may reflect underlying biological differences in the parameters. In contrast, some *HLA* allele associations reported in the literature were not supported in our population-based study using either measure. Our data are consistent with reported protective effects for the *HLA-B*27* and *HLA-C*14* alleles[23,24] and accelerated disease progression for the *HLA-B*35*, *HLA-C*04*, *HLA-C*16* and *HLA-A*29* alleles[19,23,24], but none of these associations were significant in our cohort. Using the reduced two-digit summary of *HLA* alleles, there were 66 distinct *HLA* class I alleles and 54 distinct *HLA* class II alleles in our data set, 240 exploratory comparisons were made to seek new relationships. Trends in our data were not statistically significant given the number of multiple tests, but the *HLA-C*06* and *HLA-C*18* alleles had a tendency to be associated with lower viral load, and the *HLA-C*15* and *HLA-DPB1*13* alleles were associated with slower CD4+ T-cell decline. The *HLA-C*06* and *HLA-C*18* alleles are in linkage disequilibrium with *HLA-B*5701* and *HLA-B*5702/3*, respectively[25], and the low viral load association for these two *HLA-C* alleles was dependent on the presence of *HLA-B*57*. Previously reported HLA associations with HIV transmission were not confirmed in our population-based study[18], and stratifying the men by their level of at-risk sexual activity[26] or taking the number of HIV-free days during the study into account did not reveal additional associations (data not shown). The two most noteworthy trends related to transmission (not significant when corrected for multiple tests) were an enrichment among HIV-infected men of the *HLA-C*17* allele (35 men, all *HLA-C*1701*; uncorrected Fisher's exact test *P* value = 0.02, relative risk = 2.2 with 95% confidence interval 1.1–4.7) and the *HLA-DRB1*14* allele (61 men, 82% *HLA-DRB1*1401*; uncorrected *P* value = 0.004, relative risk = 2.2 with 95% confidence interval 1.3–3.8).

Homozygosity of *HLA* loci and disease progression

HLA diversity increases the potential number of epitopes recognized, and therefore enhances the breadth of the cellular immune response. We confirmed the detrimental effect of homozygosity for *HLA* class I alleles using the viral load and the full four-digit high-resolution HLA class I classifications (Kendall's tau = 0.076, $P = 0.006$)[19]. In particular, survival analysis indicated that individuals with two homozygous class I alleles progressed more rapidly to AIDS (mean = 2,105 with 95% confidence interval, 1,425–2,782 d) compared with individuals who were heterozygous at all three class I loci (mean = 3,041 with 95% confidence interval, 2,837–3,234 d). We found no statistical support for a heterozygous advantage based on *HLA* class II alleles (Kendall's tau = 0.019, $P = 0.27$).

Association of HLA class I supertypes with disease

Next, we grouped the *HLA-A* and *HLA-B* alleles into nine discrete HLA supertypes[22] defined by their peptide-binding specificities. This biologically plausible classification scheme enabled us to reduce the 41 *HLA-A* alleles to four A supertypes and the 85 *HLA-B* alleles to five B supertypes for statistical comparisons with viral load (**Table 2a**). Three of the nine supertypes, all B supertypes, were associated with viral load. Individuals with B58s and B27s had significantly lower ($P < 0.000003$ and $P < 0.0006$, respectively) and individuals with B7s had significantly higher ($P = 0.03$) viral load distributions. Two of these associations, B7s and B27s, remained highly significant ($P = 0.001$ and $P = 0.02$, respectively) after a conservative Bonferroni correction for multiple comparisons, and the third was reduced to a trend (**Table 2b**).

We then assessed whether these supertype associations were influenced by the presence of a single *HLA* allele that was highly associated with progression of the disease[18,19,23,24]. We excluded the *HLA-B*27* allele from B27s, the *HLA-B*57* allele from B58s and both the *HLA-B*35* and *HLA-B*56* alleles from B7s. The association of the remaining B27s and B7s alleles with viral load was independent of the contribution of a particular *HLA* class I allele (**Table 2b**). Although others have found *HLA-B*27* to be associated with protection, in our study it was not; removing the *HLA-B*27* allele from the B27s supertype actually slightly improved viral load in the B27s supertype category. There was a loss of significance for B58s after removing individuals with *HLA-B*57*, but the trend was in the expected direction. The effect of *HLA-B*57* exclusion from B58s might be attributed to the loss of statistical power because of the high frequency of this allele within the B58s supertype, but we cannot exclude a potential effect of other polymorphic amino acid residues within the peptide-binding cleft on the distinct behavior of *HLA-B*57*.

HLA-supertype genotypes are highly predictive of viral load

We next developed a genotypic classification scheme to examine the influence of combinations of B supertypes on viral load. We classified B7s as high (H), B44s and B62s as medium (M), and B27s and B58s as low (L) viral load predictors. Together, these form six B-supertype genotypes (HH, HM, MM, HL, ML and LL). Unlike the associations with any specific *HLA* allele, this classification scheme describes most of the MACS cohort (352 of 481 men, or 73%). **Figure 1** shows that B-supertype combinations were highly predictive of viral load (Kendall's tau = 0.177, $P = 4 \times 10^{-7}$). To exclude confounding issues that might result from different racial backgrounds, the analysis was restricted to the 282 white males and repeated; the result remained highly significant (Kendall's tau = 0.201, $P = 2 \times 10^{-7}$). The six B-supertype genotypes were also correlated with rates of $CD4^+$ T-cell decline ($P = 0.01$).

Table 2 HLA supertype association with viral load and influence of individual alleles on the associations

a

Supertype	Allele *f*	*n*	(–)	(+)	Wilcoxon *P* value
A24s	0.1565	313	12,831	13,995	0.80
A1s	0.2045	409	14,657	10,909	0.12
A3s	0.242	484	13,611	12,596	0.96
A2s	0.2975	595	11,466	15,067	0.18
Other	0.0995	199			
B58s	0.0705	141	15,394	4,617	0.000003
B62s	0.0865	173	13,148	13,148	0.80
B27s	0.1385	277	14,473	9,783	0.03
B44s	0.264	528	11,971	14,695	0.21
B7s	0.294	588	10,260	17,234	0.0006
Other	0.14865	293			

b

Individuals carrying B7s compared with those who do not

	n	Geo. mean (s.e.m.)	*P* value
All men with B7s	230	17,234 (9,867–30,101)	0.0006
B7s, excluding *B*56*	277	16,856 (9,645–29,456)	0.001
B7s, excluding *B*56*, *B*35*	189	17,356 (10,050–29,971)	0.001
No B7s	251	10,260 (5,882–17,897)	—

Individuals carrying B27s compared with those who do not

	n	Geo. mean (s.e.m.)	*P* value
All men with B27s	118	9,783 (5,511–17,367)	0.03
B27s, excluding *B*27*	107	9,236 (5,142–16,593)	0.02
No B27s	363	14,474 (8,297–25,251)	—

Individuals carrying B58s compared with those who do not

	n	Geo. mean (s.e.m.)	*P* value
All men with B58s	63	4,617 (2,468–8,634)	0.000003
B58, excluding *B*57*	20	8,116 (4,193–15,707)	0.2531
No B58s	418	15,394 (9,002–26,325)	—

(**a**) HLA supertype associations with viral load. Listed for each supertype is the supertype 'allele' frequency (*f*) based on 996 men enrolled in the Chicago MACS; the number (*n*) out of 481 HIV-1-infected individuals with viral set points that carried one or two copies of the supertype; the geometric mean of the viral load for those who carry (+) or do not carry (–) the supertype; and the Wilcoxon rank test *P* value comparing the two distributions. (**b**) Detailed examination of the influence of single genetic allotypes on the supertypes significantly associated with viral load. Shown are the number of individuals (*n*) that carry the supertype is given; the number remaining when allotypes independently associated with progression were excluded from the set; the geometric (Geo.) mean and s.e.m. of the viral load; and a Wilcoxon *P* value comparing the supertype-carrying individuals (or subsets) with those who do not carry the supertypes.

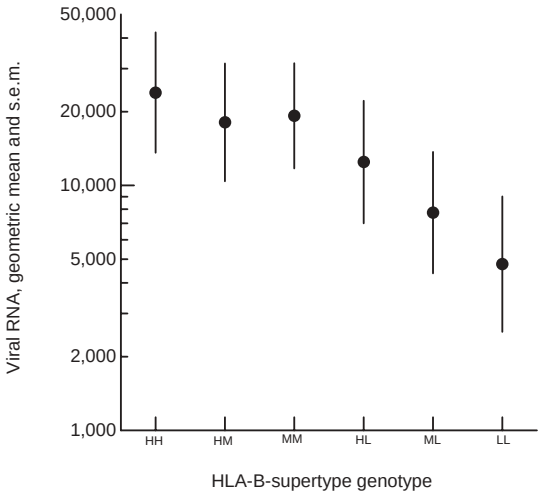

Figure 1 HLA class I B-supertype genotypes and viral load. H, B7s (correlated with high viral load); M, combination of B44s and B58s (not independently correlated with viral load); L, combination of B58s and B27s (correlated with low viral load). Combinations of H, L and M represent the person's diploid genotype. The correlation between the person's genotype and viral load is highly statistically significant ($n = 352$, Kendall's tau = 0.177, $P = 4 \times 10^{-7}$). When only white males were considered, the result remained highly statistically significant ($n = 282$, Kendall's tau = 0.202, $P = 2 \times 10^{-7}$).

HLA A- and B-supertype combinations

We used the information theoretic criterion for model selection by minimum description length (MDL)[27–29] to split the HLA supertype data into sets of genotypes that are predictive of viral load. The MDL criterion chooses the most parsimonious model to describe the data: it minimizes the sum of the lengths (in bits) of the description of the model and the data encoded by the model[27–29], in this case using a log normal model for viral load. MDL provides a strategy for determining whether we should consider all individuals as one group, or whether we are justified in assigning individuals to different groups based on their supertype combinations and viral load.

By iteratively enumerating all possible partitions of supertypes into k groups, computing the resulting description length and taking as optimum the grouping with lowest description length L_{k^*}, the optimal groupings were found to correspond to $k = 2$ (**Table 3**). Thus, the application of the MDL criterion justified splitting A- and B-supertype combinations each into two groups, one with high viral loads and one with low. Combining them into one group or splitting them into three groups yielded greater description lengths. The differences in groups were significant (Wilcoxon rank sum $P < 0.003$ for *HLA-A* and $P < 2 \times 10^{-6}$ for *HLA-B*). The inclusion of B58s-B62s individuals in the higher viral load group may be an artifact because this is a very rare combination, occurring only twice in this data set. Combinations of A supertypes that were beneficial had an average two-fold reduction in viral load relative to the high viral load group, whereas beneficial B-supertype combinations averaged a 3.5-fold reduction.

This analysis reveals new details. First, consistent with the results in **Table 2**, the most profound relationship between HLA supertype and viral load is for the *HLA-B* locus, as splitting the B supertype data yields the most substantial drop in the description length (**Table 3**). Second, a split based on A supertypes is also justified by MDL analysis. Third, we can discern a hierarchy of relationships in the groups listed in **Table 3**. For example, the B7s disadvantage is countered by having B58s, but otherwise carrying B7s relegates an individual to the high viral load category, and B58s to the low viral load category irrespective of the other allele.

HLA anchor motif frequencies in HIV proteins

One hypothesis to explain the association of HLA supertypes with viral load is that some allotypes may simply enable the recognition of more epitopes within the HIV genome, and therefore allow for greater breadth of response. A tally of the known epitopes associated with each supertype is insufficient, however as common HLA alleles are studied more often, resulting in an inherent bias for the HIV Molecular Immunology Database (www.hiv.lanl.gov/content/immunology). As an alternative approach, we scanned the viral genome for appropriate peptide-binding motifs. We found no evidence to suggest that the association between the frequency of combinations of appropriate second-position and C-terminal anchor motifs of supertypes in the HIV B-subtype consensus sequence and high viral load was less common than the anchor motifs of supertypes associated with low viral load (data not shown)[30]. Furthermore, the variability level and frequency of amino acids

Table 3 MDL analysis of HLA-supertype genotypes

k	L	L'	Group	n	Mean	Var.	Elements in group
1	595.95	595.95	1	293	13.69	7.83	All
A-supertype genotype ($n = 10$)							
2	581.41	591.41	1	106	13.06	6.44	A1s-A2s, A1s-A3s, A3s-A3s, A24s-A24s
			2	187	14.09	7.20	A2s-A2s, A2s-A3s, A2s-A24s,A3s-A24s, A1s-A24s, A1s-A1s
B-supertype genotype ($n = 15$)							
2	564.28	579.28	1	82	12.43	7.69	B7s-B58s, B44s-B58s, B62s-B62s, B27s-B44s, B27s-B58s, B58s-B58s
			2	211	14.22	5.94	B7s-B27s, B7s-B44s, B44s-B44s, B7s-B7s, B7s-B62s, B27s-B27s, B44s-B62s B27s-B62, B58s-B62s

This analysis excluded individuals with any allele that did not fit into predefined supertypes[22] ('Other' category in **Table 2a**). k, number of sets; L, sum of description lengths of the model and data encoded by the model; L', description length plus cost of finding the best model or evaluating L from all possible splits into k groups; n, number of individuals in a group. Mean and variance (Var.) for $\log_2$ viral load are given ($\log_2$ is a natural conversion, as the description length is given in bits). HLA supertypes that fall into each group are specified.

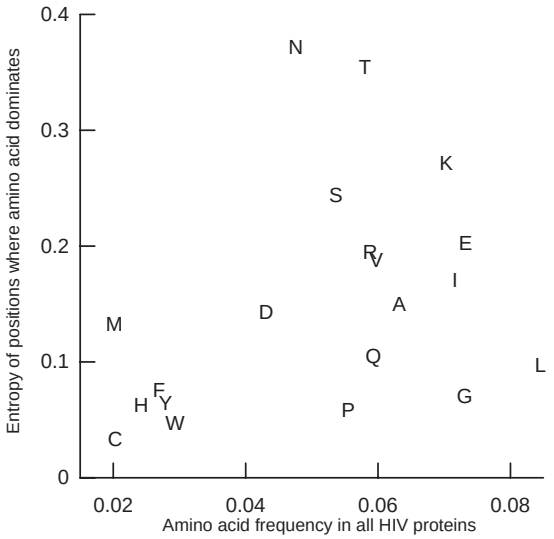

Figure 2 Amino acid frequency and entropy in HIV subtype B. Here we tested the hypothesis that HLA molecules associated with low viral loads would have either relatively common or relatively stable amino acids, or both, as anchor residues. The amino acid frequency and variability that spanned all HIV proteins was measured and compared with the anchor motif residues that define the HLA supertypes. The anchor residue in the second position is the most distinctive determinant of the B supertype, with the following specificities: P for B7s; E or D for B44s; Q, L, I, V, M or P for B62s; R, K or H for B27s; and A, S or T for B58s. B7s is associated with high viral load but P is neither rare nor variable. B27s is associated with a low viral load and favors a positively charged amino acid in the second position; R and K are common but are also quite variable.

characteristic of supertype anchor residues in HIV proteins did not reveal any clear associations with disease outcome (**Fig. 2**).

A rare allele advantage for HLA class I supertypes

A second hypothesis is that transmission of viruses with mutations that confer CTL escape occurs more frequently between partners with common HLA supertypes, which would tend to limit the breadth of their immune response.

To test this hypothesis, we compared viral loads in individuals homozygous for a supertype with the population frequency for the different A- and B-supertype alleles, and found a highly significant rare-allele advantage (Kendall's tau = 0.11, $P = 0.008$; **Fig. 3**).

Like other domestic cohort studies[31], we found that black men had lower viral loads than white men (Wilcoxon $P = 0.005$). The geometric mean for the viral load was 8,131 copies/ml (s.e.m. of 4,420–14,957 copies/ml) for the 93 black males and 14,801 copies/ml (s.e.m. of 8,629–25,386 copies/ml) for the 354 white males. The least common A (A24s) and B (B58s) supertypes were both enriched among black males compared with white males, with frequencies of 0.25 versus 0.12 for A24s and 0.11 versus 0.06 for B58s, respectively.

DISCUSSION

In this study, we have used HLA class I supertypes as a biologically plausible classification scheme for alleles with overlapping peptide-binding specificities to analyze their association with the rate of disease progression as well as the correlation of their population frequency with viral load. The most common HLA supertype, B7s, had a substantial detrimental effect on the rate of progression to AIDS. The least frequent HLA supertypes, B27s and B58s, were associated with protection against disease in our cohort. The correlation of the frequency of the HLA supertypes with viral load suggests that HIV adapts to the most frequent alleles in the population, providing a selective advantage for those patients who express rare alleles. Furthermore, the profound association of viral load with a combination of many class I *HLA-B* alleles that share a common functional trait (such as anchor residue motifs) suggests that the immune system *per se* underpins this relationship. This supports the concept that HLA molecules are crucial to the generation of a competent immune response and not a surrogate marker for other linked but unknown genes that might influence progression of the disease.

Considerable data implicate HLA-restricted, antigen-specific T-cell immunity in protection against infectious disease. In people infected with HIV, a strong CTL response is associated with the resolution of acute infection and the rates of progression of the disease[32,33]. Selection of viral mutations that result in the loss of a peptide-specific CTL response has been found in humans with acute and long-standing infection with HIV[26] and nonhuman primates infected with simian immunodeficiency virus[34]. Polymorphism within HIV reverse transcriptase may be the result of selection by CTLs[35,36]. Thus, HIV can adapt to the HLA molecule of its host by mutating the epitopes to which the CTLs respond.

Transmission of escape variants might diminish the extent of the epitopes recognized in a newly infected individual[26]. Such transmission events may occur more frequently between individuals who share the more common supertypes, so individuals that express rare alleles would have a selective advantage. Indeed, we found a strong correlation between the frequency of an HLA supertype in the

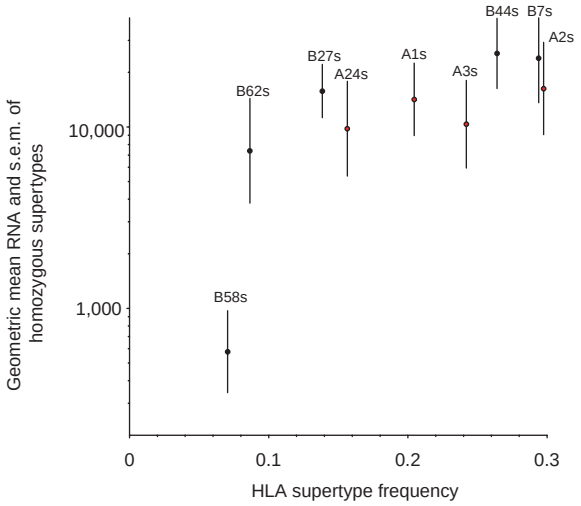

Figure 3 Correlation between HLA supertype population frequencies and viral load, indicating rare-allele advantage. Using the 220 men 'homozygous' for B supertypes, the geometric mean and s.e.m. of viral load for each supertype is plotted against the supertype frequency in our cohort; these are significantly correlated (Kendall's tau = 0.11, $P = 0.008$). Significance was maintained when we excluded the extreme B58s data points (tau = 0.078, P value = 0.05). Twenty men were homozygous for both A- and B-supertypes, so we also randomly assigned these individuals to either their A or B supertype and recomputed the test statistics ten times (average $P = 0.010$ including B58s; $P = 0.06$ excluding B58s).

population and the viral load of individuals homozygous for that supertype. This form of frequency-dependent selection, along with possible differences in viral peptide binding and presentation, may account for the association of HLA supertypes with disease progression. As other, more subtle biological factors such as differential rates of HLA class I molecule assembly and processing[37,38] may contribute to supertype associations, it will be important to establish independent validation of the rare-allele advantage in other HLA-diverse populations.

These findings are particularly notable, considering the large numbers of individuals that are needed to identify polymorphisms in natural populations that associate more frequently with a particular phenotype than expected by chance alone. By applying the MDL principle, we were able to find a hierarchy of relationships for the association of HLA supertypes with viral load (the B58s advantage seems to be dominant, for example). MDL has potential for application to other studies of complex genetic factors in relation to disease, as it provides a logical strategy for categorizing genotypes that compensates for multiple tests. As has been shown for influenza epitopes[39], and is apparent in our data, combinations of *HLA* alleles can dictate the specificity and magnitude of an immune response. Thus, a method like MDL that enables the discovery of associations among polymorphic genes that predict outcome has broad applications.

The simplest explanation for the association between HLA supertype and viral load lies in the HLA molecule's ability to bind an antigenic peptide and engage a CTL. CTLs directed against potent HIV epitopes can exert strong selective pressure on the virus[33]. Variants escape immune recognition by affecting processing, HLA binding or recognition by the T-cell receptor. Although a mutation in a crucial epitope can alter immune recognition and result in a selective advantage, there is a potential cost to viral infectivity from mutations in protein regions that are functionally or structurally constrained. This could explain the association of simian immunodeficiency virus Tat-specific escape with lower viral loads in nonhuman primates after resolution of their primary infection[31]. A virus selected by the predominant supertype in the population could compromise immune recognition of cross-reactive epitopes in individuals who share that supertype, but could function as a susceptible form in individuals that do not. Thus, the frequency of *HLA* alleles in a population that share peptide-binding specificities may drive adaptation of virus sequences in that population and help to establish the relative frequencies of amino acids in polymorphic sites.

The rare-allele advantage model of frequency-dependent selection generates several testable predictions about the interactions between a pathogen and its host. First, the model suggests that in different human populations, different *HLA* alleles would associate with the disease progression rate, and HIV disease outcome would depend at least in part on *HLA* allele frequencies and not solely on intrinsic features of the HLA molecule. Our results do not exclude the possibility that specific interactions between the HLA molecule, CTL and a particular epitope are important for AIDS risk; they only indicate that population frequencies are contributory. This finding could explain some of the differences in the HLA associations with progression of HIV disease reported previously[18]. Second, we predict that racial groups underrepresented in a panmictic epidemic population would have lower viral loads. This prediction was borne out by the lower viral loads found among the black participants in our cohort. Lastly, the data presented here suggest that epidemic pathogens have the potential to rapidly alter *HLA* allele frequencies in human populations. There is mounting evidence that escape from the CTL response profoundly influences virus evolution at the population level[36,40]. Conversely, one would expect the virus to ultimately influence host evolution in an epidemic of this magnitude in regions of the world where HIV is highly prevalent.

Our findings support a model for HLA evolution in which both overdominance and frequency-dependent selection contribute to maintaining HLA polymorphism. Still, certain *HLA* alleles may be advantageous and may have been selected because of their mediation of particularly potent CTL responses; indeed, such a selective sweep has been invoked to account for the restricted MHC class I polymorphism in the chimpanzee[41]. Our observations, however, might have been influenced by the heterogeneity of the population in our cohort; future studies may allow for more careful control and more detailed analysis than was possible in this sampling frame. Ultimately, analysis of different populations to test the predictions that arise from the model will allow us to obtain a more complete understanding of selective forces responsible for polymorphism at the HLA loci, and better understand the influence of host HLA molecules on the development of disease.

METHODS

Study participants. The study subjects were men enrolled in the Chicago component of MACS, a natural history study of men who have sex with men[30]. Participants were followed at 6-month intervals, queried about risk behaviors, tested for antibodies to HIV and had their $CD4^+$ and $CD8^+$ T-cell numbers enumerated. Infected men had antiviral therapy and levels of HIV RNA in plasma measured by quantitative RT-PCR (Roche Molecular Diagnostic Systems). All study subjects had equal access to care. Of the 1,351 men recruited for the Chicago MACS, 996 had peripheral-blood samples available for high-definition molecular *HLA* typing and 562 of these men were HIV positive. The time from seroconversion to AIDS (<200 $CD4^+$ T cells per mm^3) was established for only 64 men. Thus, we defined disease outcome either by the level of viral RNA in plasma measured after the initial burst of viral replication during acute infection (the viral level at set point, herein referred to as "viral load"; data available for 481 men) or by rate of $CD4^+$ T-cell decline over a minimum period of 2 years, including data from at least four study visits (data available for 418 men). These measurements were considered before the onset of AIDS and before the start of treatment. All men provided written informed consent according to the guidelines of the human subjects protection committee of Northwestern University.

Molecular HLA class I and class II analysis. *HLA* class I and class II loci were amplified by PCR using locus- and sequence-specific primers in a tiered typing strategy. PCR-product DNA was hybridized with established sequence-specific oligonucleotide probes and immobilized probe arrays to identify specific *HLA* class I (*HLA-A*, *HLA-B* and *HLA-C*) and class II (*HLA-DRB1*, *HLA-DQB1* and *HLA-DPB1*) alleles[42–46]. HLA haplotypes were estimated based on linkage disequilibrium patterns observed in population surveys and were informative in resolving some ambiguous combinations of alleles[25,47]. *HLA* class I and class II genotypes were expressed as four-digit alleles and subsequently reduced to either two-digit allelic groups for comparison with the previous literature, or HLA supertypes[22]. Four-digit molecular typing is necessary for supertype classification, as alleles that share a two-digit classification equivalent to serological typing and molecular typing at the serological level (particularly *HLA-B*15*) can belong to different supertypes.

Statistical analysis. Analyses were performed with the statistical packages Splus6.0 or R-project[48] or with analysis tools developed for this project. We tested the strength of the HLA association with the nonparametric Wilcoxon rank order statistic, comparing distributions of viral load or $CD4^+$ T-cell slope values in the set of people that carried the allele versus those that did not. Because our testing of previously reported associations of HLA with disease was by individual comparisons to previous findings, multiple-test corrections were thought not to be appropriate.

To test for correlations, we used the nonparametric correlation Kendall's tau statistic. It is generally implemented with an approximation to contend with ties in the data. Out of concern for the large number of ties in our data set, we also used a Monte-Carlo method to provide an unbiased P value[49]. We produced $n = 9{,}999$ surrogate data sets in which the association between the two variables was scrambled; we computed the tau statistic for each surrogate data set and for the original data, and determined the rank of the tau statistic in the sorted list of $n + 1$ tau values. If r is the rank, then the one-sided P value is given by $(n + 2 - r) \times (n + 1)$. For the two-sided test, we used the same formula but used the absolute value of the tau statistic in this procedure. While the Splus Kendall's tau and Monte Carlo did not always agree exactly, reported significant P values were significant in both the Splus and Monte Carlo tests.

There were three new hypotheses tested in this study: correlation of viral load with individual supertypes, correlation of viral load with B-supertype combinations and correlation of viral load with supertype frequency. Of these, the first involved selection from multiple tests (nine tests were done) and required a Bonferroni multiple-test correction, and the other two were independent.

We used the MDL method to determine whether the viral load data could be divided into groups with high or low log RNA values, constrained by the need to keep sets of individuals with the same HLA pairs in the same group. We used an exhaustive search over all possible partitions of the data to resolve the optimum solution shown in **Table 3**. To describe this method in terms of hypothesis testing, we tested many alternative hypotheses. The MDL technique compensates for this multiple testing by effectively subtracting from the log likelihood the term $\log_2 n_{\text{partition}}$, which corresponds to the amount of information required to specify a given partition out of all those tested. This increases the critical z-value by this same amount. The chosen partition was then corroborated by traditional hypothesis testing to test the most favorable partition, under the null hypothesis that the data are homogeneous. Even with the conservative Bonferroni correction, the null hypothesis was soundly rejected.

ACKNOWLEDGMENTS

We thank M. Vinson, C. Okoye, J. Joffee-Block and P. Otto for technical assistance, and L. Jacobson for discussions of the data. The project was funded by the National Cancer Institute (R01-HD37356 to S.W.), the National Institute of Allergy and Infectious Diseases (P30-CA79458 to S.W.) and the National Institutes of Health (U01-AI-35039 to J.P. and S.W.). Additional support was provided by the Elizabeth Glazer Pediatric AIDS Foundation (R.F. and B.K.), a Los Alamos National Laboratory Program Developmental Award (B.K., M.F. and J.T.), the My Brother Joey Foundation (E.T.) and an anonymous foundation (S.W.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 22 March; accepted 2 June 2003
Published online 22 June 2003; doi:10.1038/nm893

1. Trowsdale, J. & Campbell, R.D. Complexity in the major histocompatibility complex. *Eur. J. Immunogenet.* **19**, 45–55 (1992).
2. Bjorkman, P.J. & Parham, P. Structure, function, and diversity of class I major histocompatibility complex molecules. *Annu. Rev. Biochem.* **59**, 253–288 (1990).
3. Buus, S., Sette, A., Colon, S., Miles, C. & Grey, H.M. The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenetic peptides. *Science* **235**, 1353–1358 (1987).
4. Hill, A.V.S. *et al.* Common West African HLA antigens are associated with protection from severe malaria. *Nature* **352**, 595–600 (1991).
5. Little, A.M. & Parham, P. Polymorphism and evolution of HLA class I and II genes and molecules. *Rev. Immunogenet.* **1**, 105–123 (1999).
6. Hughes, A.L., Ota, T. & Nei, M. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.* **7**, 515–524 (1990).
7. Slatkin, M. & Muirhead, C.A. A method for estimating the intensity of overdominant selection from the distribution of allele frequencies. *Genetics* **156**, 2119–2126 (2000).
8. Bodmer, W.F. Evolutionary significance of the HLA system. *Nature* **237**, 139–145 (1972).
9. Hill, A.V.S. The immunogenetics of human infectious diseases. *Ann. Rev. Immunol.* **16**, 593–617 (1998).
10. Propato, A. *et al.* Spreading of HIV-specific CD8+ T-cell repertoire in long-term nonprogressors and its role in the control of viral load and disease activity. *Hum. Immunol.* **62**, 561–576 (2001).
11. MacDonald, K.S. *et al.* Human leucocyte antigen supertypes and immune susceptibility to HIV-1, implications for vaccine design. *Immunol. Lett.* **79**, 151–157 (2001).
12. Bertoni, R. *et al.* Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis. *J. Clin. Invest.* **100**, 503–513 (1997).
13. Tomiyama, H., Yamada, N., Komatsu, H., Hirayama, K. & Takiguchi, M. A single CTL clone can recognize a naturally processed HIV-1 epitope presented by two different HLA class I molecules. *Eur. J. Immunol.* **30**, 2521–2530 (2000).
14. Altfeld, M.A. *et al.* Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif. *J. Virol.* **75**, 1301–1311 (2001).
15. Clarke, B. The ecological genetics of host-parasite relationships. in *Genetic aspects of host-parasite relationships.* (eds. Taylor, A.E.R. & Muller, R.M.) 87–103 (Blackwell Oxford, 1976).
16. Howard, J.C. MHC organization of the rat: evolutionary considerations. in *Evolution and Vertebrate Immunity* (eds. Kelsoe, G. & Schulze, D.H.) 397–411 (University of Texas Press,, Austin, 1987).
17. Roger, M. Influence of host genes on HIV-1 disease progression. *FASEB J.* **12**, 625–632 (1998).
18. Trachtenberg, E.A. & Erlich, H.A. A review of the role of the human leukocyte antigen (HLA) system as a host immunogenetic factor influencing HIV transmission and course of infection with progression to AIDS. in *HIV Molecular Immunology Database* (eds. Korber, B. *et al.*) 1–60 (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM., 2001).
19. Carrington, M. *et al.* HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* **283**, 1748–1752 (1999).
20. Keet, I.P. *et al.* Consistent associations of HLA class I and II and transporter gene products with progression of human immunodeficiency virus type 1 infection in homosexual men. *J. Infect. Dis.* **180**, 299–309 (1999).
21. Marsh, S.G.E., Parham, P. & Barber, L.D. *The HLA FactsBook* **398** (Academic Press, New York, 2002).
22. Sette, A. & Sidney, J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* **50**, 201–112 (1999).
23. Kaslow, R.A. *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat. Med.* **2**, 405–411 (1996).
24. Hendel, H. *et al.* New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. *J. Immunol.* **162**, 6942–6946 (1999).
25. Begovich, A.B. *et al.* Polymorphism, recombination and linkage disequilibrium within the HLA class II region. *J. Immunol.* **148**, 249–258 (1992).
26. Goulder, P.J. *et al.* Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* **412**, 334–338 (2001).
27. Rissanen, J. *Stochastic Complexity in Statistical Inquiry.* **171** (World Scientific (Singapore), 1989).
28. Rissanen, J. Hypothesis selection and testing by the MDL principle. *Comput. J.* **42**, 260–269 (1999).
29. Li, M. & Vitanyi, P. *An Introduction to Kolmogorov Complexity and its Applications.* **546** (Springer-Verlag (New York), 1993).
30. Nelson, G.W., Kaslow, R. & Mann, D.L. Frequency of HLA allele-specific peptide motifs in HIV-1 proteins correlates with the allele's association with relative rates of disease progression after HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **94**, 9802–9807 (1997).
31. Anastos, K. *et al.* Association of race and gender with HIV-1 RNA levels and immunologic progression. *J. Acquir. Immune Defic. Syndr.* **24**, 218–226 (2000).
32. Phair, J. *et al.* Acquired immune deficiency syndrome occurring within 5 years of infection with human immune deficiency virus type-1: the Multicenter AIDS Cohort Study. *J. Acquir. Immune Defic. Syndr.* **5**, 490–496 (1992).
33. Saah, A.J. *et al.* Predictors of the risk of development of acquired immunodeficiency syndrome within 24 months among gay men seropositive for human immunodeficiency virus type 1: a report from the Multicenter AIDS Cohort Study. *Am. J. Epidemiol.* **135**, 1147–1155 (1992).
34. Allen, T.M. *et al. Tat*-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* **407**, 386–390 (2000).
35. van der Burg, S.H. *et al.* HIV-1 reverse transcriptase-specific CTL against conserved epitopes do not protect against progression to AIDS. *J. Immunol.* **159**, 3648–3654 (1997).
36. Moore, C.B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296**, 1439–1443 (2002).
37. Hill, A., Takiguchi, M. & McMichael, A. Different rates of HLA class I molecule assembly which are determined by amino acid sequence in the alpha 2 domain. *Immunogenetics* **37**, 95–101 (1993).
38. Williams, A., Peh, C.A. & Elliott, T. The cell biology of MHC class I antigen presentation. *Tissue Antigens* **59**, 3–17 (2002).
39. Boon, A.C. *et al.* The magnitude and specificity of influenza A virus-specific cytotoxic T-lymphocyte responses in humans is related to HLA-A and -B phenotype. *J. Virol.* **76**, 582–590 (2002).
40. Yusim, K. *et al.* Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J. Virol.* **76**, 8757–8768 (2002).

41. de Groot, N.G. *et al*. Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees. *Proc. Natl. Acad. Sci. USA* **99**, 11748–11753 (2002).
42. Bugawan, T.L., Begovich, A.B. & Erlich, H.A. Rapid HLA-DPB typing using enzymatically amplified DNA and nonradioactive sequence specific oligonucleotide probes. *Immunogenetics* **34**, 413 (1991).
43. Bugawan, T.L. & Erlich, H.A. Rapid typing of HLA-DQB1 DNA polymorphism using nonradioactive oligonucleotide probes and amplified DNA. *Immunogenetics* **33**, 163–170 (1991).
44. Bugawan, T.L., Apple, R. & Erlich, H.A. A method for typing polymorphism at the HLA-A locus using PCR amplification and immobilized oligonucleotide probes. *Tissue Antigens* **44**, 137–147 (1994).
45. Scharf, S.J., Griffith, R.L. & Erlich, H.A. Rapid typing of DNA sequence polymorphism at the HLA-DRB1 locus using the polymerase chain reaction and nonradioactive oligonucleotide probes. *Hum. Immunol.* **30**, 190–201 (1991).
46. Erlich, H.A. & Trachtenberg E.A. PCR-based methods of HLA typing. in *Molecular Epidemiology: A Practical Approach* (eds. Carrington, M. & Hoelzel, A.R.) **181– 207** (Oxford University Press, Oxford, 2001).
47. Cao, K. *et al*. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum. Immunol.* **62**, 1009–1030 (2001).
48. Venables, W.N. & Ripley, B.D. *Modern Applied Statistics with S-PLUS*. **501** (Springer (New York), 1999).
49. Hope, A.C.A. A simplified Monte-Carlo significance test procedure. *J. R. Stat. Soc. Ser. B*, 582–598 (1968).