Extensive HLA-driven viral diversity following a narrow-source HIV-1 outbreak in rural China

*Tao Dong,[1] *Yonghong Zhang,[1,2] *Ke Yi Xu,[2,3] *Huiping Yan,[2] Ian James,[4] Yanchun Peng,[1] Marie-Eve Blais,[1] Silvana Gaudieri,[4,5] Xinyue Chen,[2] Wenhui Lun,[3] Hao Wu,[2] Wen Yan Qu,[3] Tim Rostron,[1] Ning Li,[2] Yu Mao,[3] Simon Mallal,[4] Xiaoning Xu,[1] Andrew McMichael,[1] *Mina John,[4] and *Sarah L. Rowland-Jones[1]

[1]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, United Kingdom; [2]Beijing You An Hospital, Capital Medical University, Beijing, People's Republic of China; [3]Beijing Ditan Hospital, Capital Medical University, Beijing, People's Republic of China; [4]Centre for Clinical Immunology and Biomedical Statistics, Institute for Immunology and Infectious Diseases, Murdoch University and Royal Perth Hospital, Perth, Australia; and [5]School of Anatomy and Human Biology and Centre for Forensic Science, University of Western Australia, Perth, Australia

Obstacles to developing an HIV-1 vaccine include extensive viral diversity and lack of correlates of protective immunity. High mutation rates allow HIV-1 to adapt rapidly to selective forces such as antiretroviral therapy and immune pressure, including HIV-1–specific CTLs that select viral variants which escape T-cell recognition. Multiple factors contribute to HIV-1 diversity, making it difficult to disentangle the contribution of CTL selection without using complex analytical approaches. We describe an HIV-1 outbreak in 231 former plasma donors in China, where a narrow-source virus that had contaminated the donation system was apparently transmitted to many persons contemporaneously. The genetic divergence now evident in these subjects should uniquely reveal how much viral diversity at the population level is solely attributable to host factors. We found significant correlations between pairwise divergence of viral sequences and HLA class I genotypes across epitope-length windows in HIV-1 Gag, reverse transcriptase, integrase, and Nef, corresponding to sites of 140 HLA class I allele-associated viral polymorphisms. Of all polymorphic sites across these 4 proteins, 24%-56% were sites of HLA-associated selection. These data confirm that CTL pressure has a major effect on interhost HIV-1 viral diversity and probably represents a key element of viral control. (*Blood*. 2011;118(1):98-106)

Introduction

Despite > 2 decades of research, the critical components of protective immunity to HIV-1 infection remain inadequately defined. The general consensus that CD8[+] CTLs provide a major force controlling viral replication[1] was challenged by the failure of a Merck adenovirus 5 recombinant candidate HIV-1 vaccine to confer protection despite inducing virus-specific CTLs in most recipients.[2] The halting of the STEP vaccine trial prompted calls for a fundamental reevaluation of the role of the different elements of the immune response to HIV-1 infection.

One approach to defining the effect of cellular immune responses on viral control is to determine the extent to which virus evolution is dictated by HLA class I–restricted T-cell targeting of particular viral epitopes. HIV-1 undergoes diversification in an infected person over the course of disease,[3] leading to the coexistence of multiple "quasispecies." The high rate of mutation is largely because of the error-prone nature of reverse transcriptase and allows the virus to respond rapidly to selection pressure from forces such as antiretroviral therapy (ART) and the host immune response.[4] The accumulation of viral variants in the infected person is reflected in the extraordinary diversity of circulating viruses, even within viral subtypes, shown in population-based studies.[5] Virus-specific CTLs constitute an important selective force on viral evolution: viral escape from CTL pressure in the infected person is well described[6,7] and tends to follow stereotypic mutational pathways on the basis of the HLA-restriction of CTL

epitopes, just as ART resistance mutations are characteristic for particular drugs.[8] The extent of CTL selection at the population level was first shown by Moore et al,[9] who analyzed the frequency of amino acid substitutions departing from a population consensus reverse transcriptase (RT) sequence in HIV-infected persons in an Australian population as a function of their HLA-A or -B genotypes. The analytical approach investigated the correlation between individual HLA types and autologous HIV-1 RT polymorphisms and used multivariate methods to adjust for coinheritance of HLA alleles within the MHC, as well as covarying codons in RT. This adjustment aimed to distinguish associations that arose directly from viral escape mutation within HLA-restricted epitopes from those caused indirectly by linked HLA alleles, compensatory viral mutations, or subtype-specific viral polymorphisms. This study concluded that CTL selective pressure makes a major contribution to viral intrahost diversity and, in some cases, drives fixation of HLA-adapted residues in the population, implying that certain HIV-1 epitopes may become less immunogenic over time.[9-11] Subsequent methods were developed that used phylogenetic trees to impute shared viral lineage between HIV sequences and to adjust explicitly for "founder effects," in which viruses related by common lineage and also enriched in immunogenetically distinct subpopulations may lead to correlations between certain HLA types and viral polymorphisms.[12] With the use of these methods, it was argued that the extent of CTL selection in viral evolution may

have been overestimated.[12] Another study used mathematical modeling to estimate the contribution of CTLs to driving HIV-1 variation in chronic[13] infection and also concluded that selection by CTLs plays only a minor role[14] (although it could be argued that this study may have underestimated the selection imposed by the potent CTL response in acute HIV-1 infection[15-17]). Additional support for the concept that at the population level certain HIV-1 epitopes would become less immunogenic over time was provided by the work of Scherer et al.[13] Large population-based studies in geographically diverse populations that used several methods of phylogenetic correction have shown extensive HLA allele–specific polymorphism across the HIV-1 subtype B and C proteomes[5,18,19]: these methods require large sample sizes for adequate statistical power. In addition, the methods are based on the principle that CTL selection could not itself drive any phylogenetic similarity between viral sequences among subpopulations with many shared HLA alleles, although one study has suggested that phylogenetic clustering at more terminal branches in a tree, such as within viral subtypes, could be influenced by immune selection.[20] All of these issues make statistical estimations of CTL selection in driving HIV-1 variation a challenging undertaking in most HIV-1 epidemics, particularly those with complex subtype admixtures, strong host population substructures, or complex viral transmission networks.

Here, we describe the unusual situation of a large population-based outbreak of HIV-1 infection occurring in an isolated rural community in Henan province in central China after participation in a paid plasma donation scheme in the village. Such schemes operated in various parts of Henan and surrounding provinces between 1980 (at the earliest) and 1996; however, donations within this community (referred to as "SM village") only occurred within a relatively narrow period between 1993 and 1995. It is thought that HIV-1 transmissions among paid plasma donors in China occurred as a result of contamination of blood collection equipment or pooled red cells being returned to donors[21]: a previous study of the *p17* region of *gag* and C2-V3 region of e*nv,* which included 89 persons sampled across 15 other Henan communities, suggested that the paid plasma donation/blood transfusion-associated HIV-1 subtype B′ epidemic in China is monophyletic.[22] We were able to ascertain fully all surviving HIV-infected persons in SM village, based on community-based HIV screening programs undertaken in 2004-2005, and to establish epidemiologically that HIV-1 infection probably occurred by the same route and in the same timeframe in all study subjects. We present analysis of HIV-1 *gag, pol*, and *nef* proviral sequences and HLA class I genotypes from 231 surviving HIV-1 infected plasma donors in SM village, derived from samples collected ~ 10-12 years after primary infection and (in most cases) before ART exposure. Because of the unique epidemiologic characteristics of this outbreak, which suggest contemporaneous infection of multiple hosts from an unusually narrow source by the same route of infection, we analyzed the viral sequence data to determine whether this was apparent in their phylogenetic relationships. We then sought to determine the extent to which HLA-related selection pressure driving intrahost viral evolution during the decade since infection accounted for interhost HIV-1 diversity and evolution.

Methods

Ethics statement

Ethical approval was obtained from Beijing Youan Hospital and the University of Oxford Tropical Ethics Committee (OXTREC).

HIV-1 sequencing

HIV-1 *gag, pol* (RT and integrase regions), and *nef* were amplified by nested PCR from proviral DNA, and bulk sequences were derived as described previously.[23]

HLA genotyping

Low-resolution (2-digit) HLA class I molecular typing was performed with an Amplification Refractory Mutation System with sequence-specific primers at the Human Immunology Unit, Weatherall Institute of Molecular Medicine, Oxford. Deviations from Hardy-Weinberg equilibrium were tested with the Arlequin v3.1 software.[24]

Multiple comparisons

False discovery rates and associated q-values[25] need to take account of both the discreteness of the test statistics and strong correlations between tests. We obtained the null *P* value distributions by replicating the analysis to create the appropriate tables and marginal frequencies, fixing the margins but imputing random hypergeometric table values subject to these fixed margins. Because of the replication of similar tables with corresponding marginal frequencies within each analysis, 50 imputed random tables were sufficient to estimate the null distribution. False discovery rates and q-values were then obtained by comparing the observed and null *P* value distributions.[26]

HIV-1 phylogenetic analysis

The HIV-1 phylogenetic analysis is described in supplemental Methods (available on the *Blood* Web site; see the Supplemental Materials link at the top of the online article).

Phylogenetic stratification of HLA allele-HIV-1 polymorphism associations

The phylogenetic stratification of HLA allele-HIV-1 polymorphism associations is described in supplemental Methods.

T-cell assays

18mer peptides that contained residues with strong HLA associations were used in ELISPOT assays with the use of PBMCs derived from donors with the relevant HLA type who had not yet developed the HLA-associated mutation in vivo. CTL lines/clones were generated as described previously.[23] Optimal epitope peptides and HLA restriction were determined with T-cell clones tested against truncated peptides and B-cell lymphoblastoid cell lines with matching single HLA class I molecules, as previously described.[27]

Results

Study population

Samples were collected from all identified former plasma donors with chronic HIV-1 infection, living in SM village, Henan province, China. SM village is a close-knit and geographically isolated rural community, in which most local residents have lived for several generations and have intermarried between families. Between 1993 and 1995, many residents in this village joined a scheme for paid plasma donation, many of whom donated their plasma repeatedly. Most members of the cohort were not aware that they had been infected with HIV-1 until 2004 when large-scale HIV screening programs were initiated in China. We estimate that 407 former plasma donors in SM village acquired HIV infection, based on the identification of 258 HIV-1–infected adults in 2005 and reports of 149 premature adult deaths with symptoms compatible with HIV-1 disease before 2004. HIV-1 infection was not

detected in persons residing in the village during 1993-1995 that did not donate plasma, suggesting that infection was not easily transmissible through the village by routes not associated with plasma donation. Of the surviving HIV-1–infected patients, 258 were recruited into this study; none was treated with ART before 2004. Viral sequence data were generated from all 258 subjects (using samples obtained between 2005 and 2007), and HLA typing was completed for 231 of these patients. The epidemiologic data suggest that all the cohort members probably acquired HIV-1 infection by the same route during the same time period and subsequently progressed to diverse disease outcomes without ART for the first 9-10 years of infection. A total of 89 subjects received ART (for various lengths of time) in the 1-3 years before samples were obtained for viral sequencing.

HLA class I allele distribution

Two-digit HLA typing showed a hierarchical structure for HLA-A alleles, dominated by HLA-A*02 (30%), HLA-A*11 (13.4%), A*24 (14.7%), and A*33 (10.7%). HLA-B*40 was the most prevalent B allele (14.5%), whereas HLA-B*51 and B*13 were observed at frequencies of 10.8% and 10.3%, respectively. Among HLA-C types, HLA-Cw*03 was the dominant allele (20%), followed by HLA-Cw*07 (16.1%), Cw*06 (14.9%), Cw*08 (12.4%), and Cw*01 (11%). This HLA distribution is generally similar to that reported in other Han Chinese cohorts.[28,29] The rate of heterozygosity for HLA-A and HLA-C alleles did not suggest any deviation from the Hardy-Weinberg equilibrium; however, for HLA-B alleles analyzed separately, the observed rate of heterozygosity was less than expected (0.86 observed vs 0.93 expected; $P = .025$, SD = 0.000 10).[30] Associations with low viral load that remained significant after correction for multiple comparisons were observed in persons with carriage of HLA-A*30 and HLA-B*51 (data not shown).

HIV-1 sequence diversity in the SM cohort is consistent with an outbreak from closely related strains

We constructed maximum likelihood phylogenetic trees of SM cohort HIV-1 *gag*, *pol*, and *nef* sequences. SM cohort sequences across all 3 proteins clustered with the subtype B′ reference sequence YN.RL42 as well as sequences obtained from GenBank derived from plasma donation associated infections in neighboring regions.[31] In particular SM cohort p17 *gag* regions interspersed with matched length subtype B′ p17 sequences derived from paid plasma donors from neighboring cities in Henan province, examined in a previous study by Zhang et al[22] (Figure 1A). In contrast, p17 sequences from intravenous drug users from 3 different regions of southern and western China (subtypes CRF07 and CRF08) and those with probable sexually acquired infection from Beijing (subtype B and recombinants) examined previously[22] clustered separately from each other and from SM cohort sequences. As noted by Zhang al,[22] there was no apparent clustering by geographic location (ie, clustering among SM cohort sequences and other plasma donation-related sequences from outside SM village); rather, there was clear clustering on the basis of route of transmission across all these Chinese populations.

The phylogenetic patterns reflected the genetic distance evident within and between groups of sequences. The mean genetic distance within the SM cohort was comparable to that previously observed among Henan plasma donation-associated sequences[22,32] (6% vs 4.4% in p17, respectively) and was only 3% when *gag, pol,* and *nef* were considered, suggesting a restricted diversity, similar to
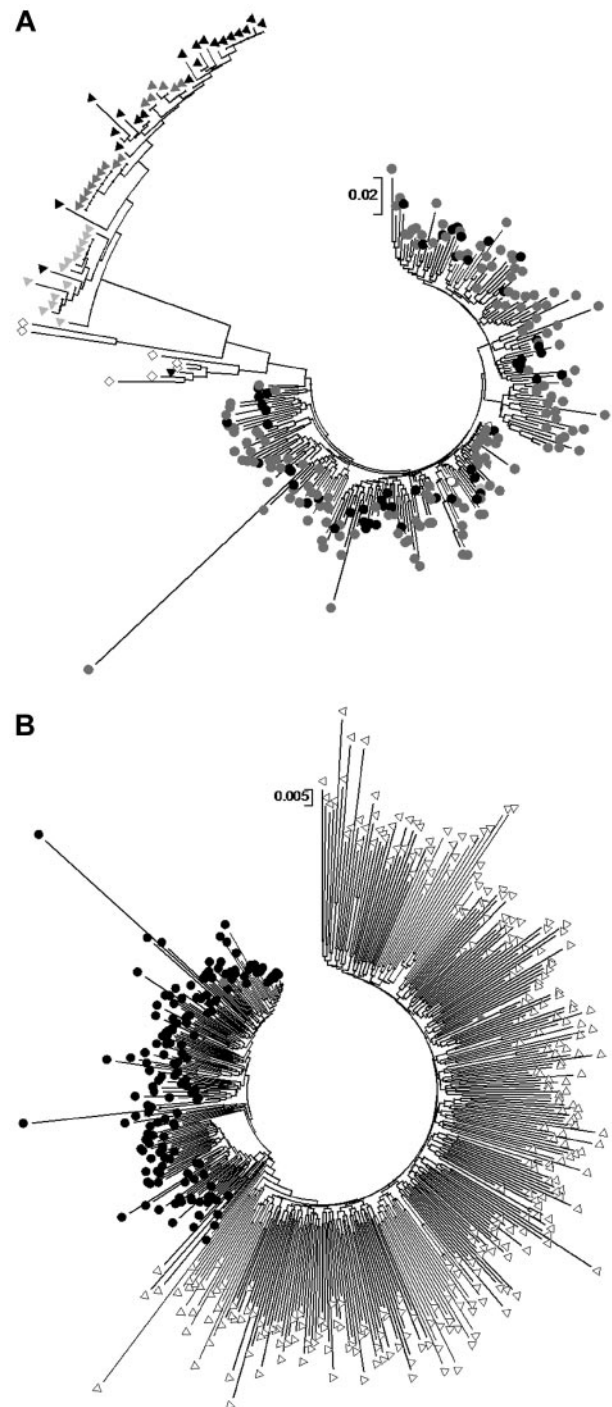

Figure 1. Maximum likelihood phylogenetic trees and full-length *gag* sequences. (A) Maximum likelihood phylogenetic trees of SM cohort *p17* sequences (black circles) shown with length-matched publicly available sequences derived from plasma donation-associated HIV-1 infection from other cities in Henan as described in Zhang et al[22] (gray circles), a subtype B′ reference sequence (open circle), injecting drug user–associated *p17* sequences also generated[20] (triangles of different colors from 3 different regions in China) and sexual-transmission sequences from Beijing (open diamonds). (B) Full-length *gag* sequences from SM cohort subjects (black circles) are shown in a maximum likelihood phylogenetic tree with matched-length gag sequences sampled from a subtype B-infected population in the United States (open triangles). Because of the sample size, a bootstrap value from 500 replications was only obtained for the *nef* maximum likelihood tree and was found to be 87% for the SM cluster. We obtained bootstrap values for gag and pol clusters using neighbor-joining trees, which shared the same topology as maximum likelihood trees, and these were both > 80%.
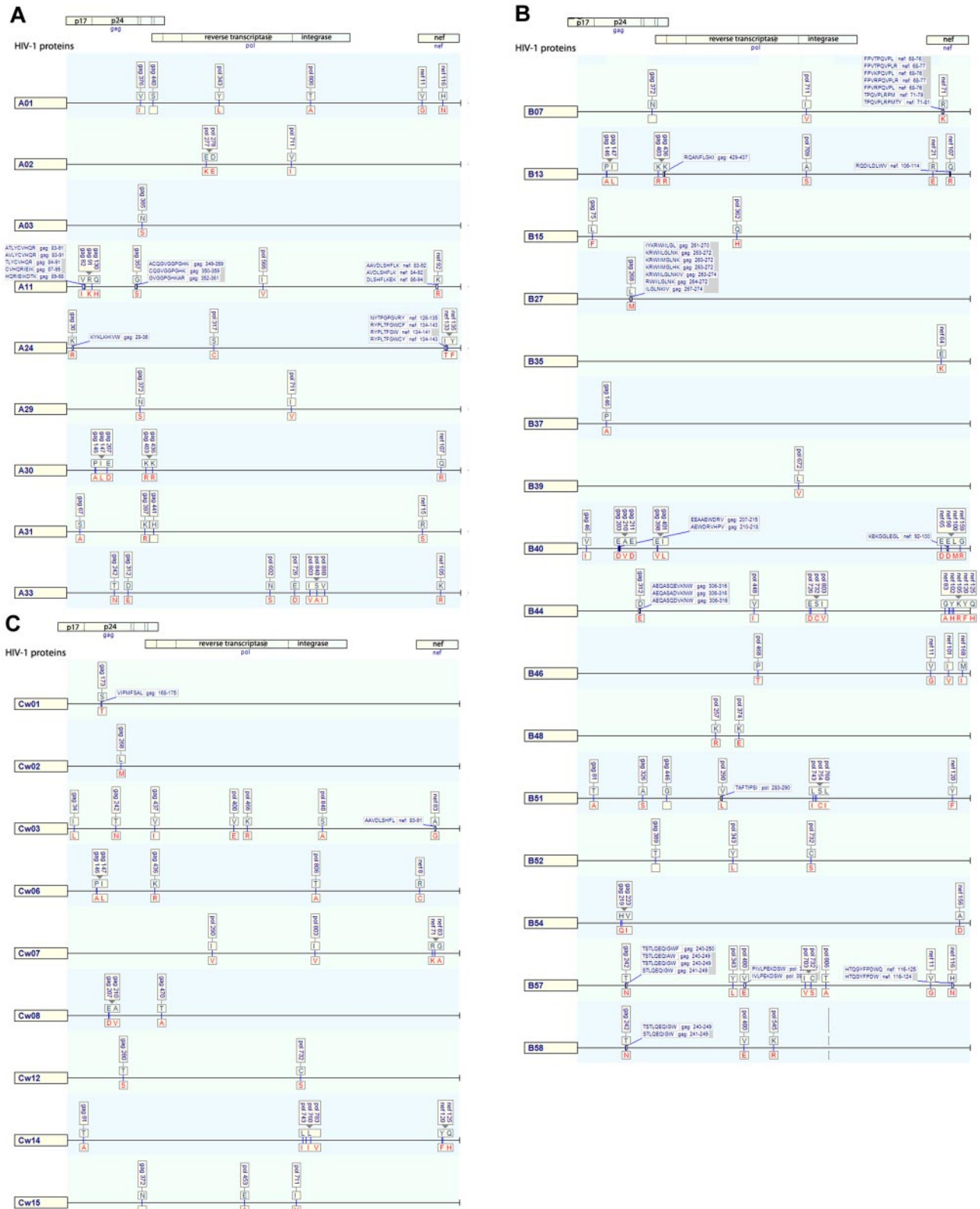
Figure 2. Maps of unique HLA-associated adaptations in HIV-1 Gag, Pol, and Nef. Maps of unique HLA-associated adaptations (q-value ≤ 0.2) in HIV-1 Gag, Pol, and Nef, grouped for HLA-A alleles (A), -B alleles (B), and -C alleles (C). The nonadapted (susceptible/revertant) amino acids are displayed above the line in blue text and adapted amino acid are below the line in red text. Locations of published CD8 T-cell epitopes are shown as boxed labels at association sites.

that seen occurring within an infected person over time or between transmission pairs.

We also compared the SM cohort HIV-1 *gag, pol*, and *nef* sequences with matched length segments of HIV-1 derived from a large population-based cohort in the United States with respect to genetic distances and phylogenetic relationships (*gag* tree shown in Figure 1B). Although the US sequences are subtype B, they derive from a large complex, long-standing epidemic[32] in which the

predominant mode of transmission is sexual, presumably with multiple sources of viral ingress into and multiple networks of transmission within the population. As expected, SM cohort sequences clustered separately across all 3 genes examined with strong bootstrap support. Genetic distances calculated with full *gag* sequences indicated that the average distance within the US cohort was 8% compared with 3% in the SM cohort, and the mean distance between them was 7%. In addition, average polymorphism rates and entropies over matched segments of *gag* in sequences drawn from the US cohort and the SM cohort were compared. The average entropy over 461 positions was 0.136 ± 0.214 in the SM cohort and 0.193 ± 0.299 in the US cohort. The average polymorphism rates were 0.039 ± 0.079 and 0.059 ± 0.110 in the SM and US populations, respectively. By both diversity measures, the SM cohort sequences had approximately one-half the level of diversity of that seen in the comparator multifounder cohort, again consistent with a narrow source epidemic. These data show the extent to which population diversity can be driven by within-patient sequence evolution alone even in a population with a relatively restricted genetic (including HLA) repertoire.

Finally, to provide further supportive evidence of the route of transmission, we sought to estimate the age of the SM cohort cluster with the use of an established Bayesian Markov Chain Monte Carlo approach, as implemented in the program BEAST v1.5, with length of chain of 30 million and previously reported substitution rate for HIV-1 subtype B *pol*.[33] This analysis indicated that the SM cohort sequences had an estimated time to most recent common ancestor of 15.01 mean years, with 95% confidence interval between 10.7 and 19.8 years, which accommodates the known period of plasma donation in SM village and suggests that no infections in this cohort occurred more recently than 1995, when plasma donation ended in the village.

HLA-HIV polymorphism associations at the population level

Although a rapidly dispersing, narrow source outbreak should not, by definition, be subject to within-cohort founder effects in the computation of HLA-HIV-1 polymorphism associations, we used a published method for computing associations which still incorporates viral sequence relatedness.[34] We detected a total of 141 statistically significant associations between HLA-A (28.4%), HLA-B (48.9%), and HLA-C (22.7%) alleles and divergence from the population consensus amino acid at single amino acid residues within HIV Gag, RT, integrase, and Nef with *P* values at or below the cutoff at which a 20% false positive rate (q-value ≤ 0.2) would be expected. All but one of these retained significance after adjustment for sequence clustering, consistent with a narrow source epidemic without strong founder effects. The final 140 associations were then plotted in HLA allele–specific maps to indicate their distribution, most probable amino acid substitution, and relationship to published CTL epitopes with a matching HLA restriction (Figure 2; supplemental Table 1). The most intense HLA-associated selection was observed in Nef (number of HLA associations per codon, 0.165), followed by Gag (0.112), integrase (0.079), and RT (0.05).

Because some subjects had received ART before providing samples for viral sequencing, we investigated whether potential ART mutations in the *pol* sequences that we analyzed could have confounded our analysis. We noted that 2 HLA class I–associated polymorphisms in RT coincided with known ART resistance mutations, namely pol 343 Y-L (Y188L), a non–nucleoside RT inhibitor resistance mutation that was associated with HLA-B57

and A1 in our cohort, and pol 374 K-E (K219E), a nucleoside RT inhibitor resistance mutation that was linked with HLA-B48. However, when we compared the frequency of mutations between ART-treated patients and the cohort overall, we saw much higher frequencies of the mutations in subjects with the relevant HLA allele than in treated persons (data not shown); therefore, we conclude that HLA class I alleles represent the main selective force for these mutations rather than drug resistance.

Contribution of intrahost HLA-associated selection to interhost HIV diversity

Given that epidemiologic history, genetic distance data, phylogenetic patterns, and the HLA associations analysis were consistent with a narrow source epidemic, the genetic distance between viral sequences within the cohort should reflect sequence evolution within individuals, to some extent after viral adaptation to each host's HLA-restricted CTL responses. The HLA associations at single residues indicate the individual changes driven by HLA-associated selection. At sites with ≥ 5 persons in the population with a nonconsensus amino acid present, and counting only the phylogeny-adjusted HLA associations with q-value < 0.2, the proportion of polymorphic sites subject to HLA-associated change was 30% in Gag, 56% in integrase, 24% in RT, and 32% in Nef.

We further hypothesized that the overall contribution of immune selection to viral diversity as a whole (beyond single residues) could also be determined by testing for the significance of correlations between HLA allele matching and similarity in viral sequences on a pair-wise basis. For each pair of persons we calculated a dissimilarity score between their viral sequences on the basis of amino acid nonagreements and a corresponding score on the basis of their HLA-A/B/C allele matching. We then looked for correlations between these scores over Gag, RT, integrase, and Nef sequences. Correlations were tested over the full-length proteins and then localized correlations were tested over sliding intervals of 10 residues across each protein, representing an approximate "epitope-length" window. All analyses were performed with Tibco Spotfire S+8.1 (Tibco Software Inc). Because the dissimilarity scores were not independent across all pairs of persons, significance was assessed by randomization tests in which HLA genotypes and viral sequences were permuted, and the standard R^2 was compared with the randomization distribution. Permutations (n = 500) were used to estimate *P* values, thus truncating *P* values at 1/500. Divergences in viral sequences on the basis of full protein length were not significantly associated with HLA mismatching for Gag, RT, integrase, or Nef; however, when shorter sliding intervals of 10 amino acids were considered, the localized correlations between HLA and sequence dissimilarities became significant across Gag, RT, integrase, and Nef (Figure 3). Many of the regions of strong HLA-viral correlation correspond to the HLA associations computed at single residues with the use of alternative methods (Figure 2); however, here it is the overall CD8 T-cell influence on pairwise diversity, rather than individual HLA allele-associated substitutions that are made evident. Notably, there are strong peaks of significance in Nef corresponding to the HLA-A24–associated change at position 135 and in integrase corresponding to HLA-A33–, -B58–, and -Cw3–associated substitutions at position 125. The observation that strong HLA correlations with viral divergence are only apparent in localized windows is consistent with the immune system's "view" of HIV, not as whole functional proteins or virus, but as a collection of short peptide lengths. Within this geography of "immunologically relevant" sequence windows,

Figure 3. Plots of significance (-logP) of correlations between nonmatching in HLA-A, -B, and -C genotypes combined and viral sequence dissimilarity over sliding windows of 10 amino acids in each HIV-1 protein examined.
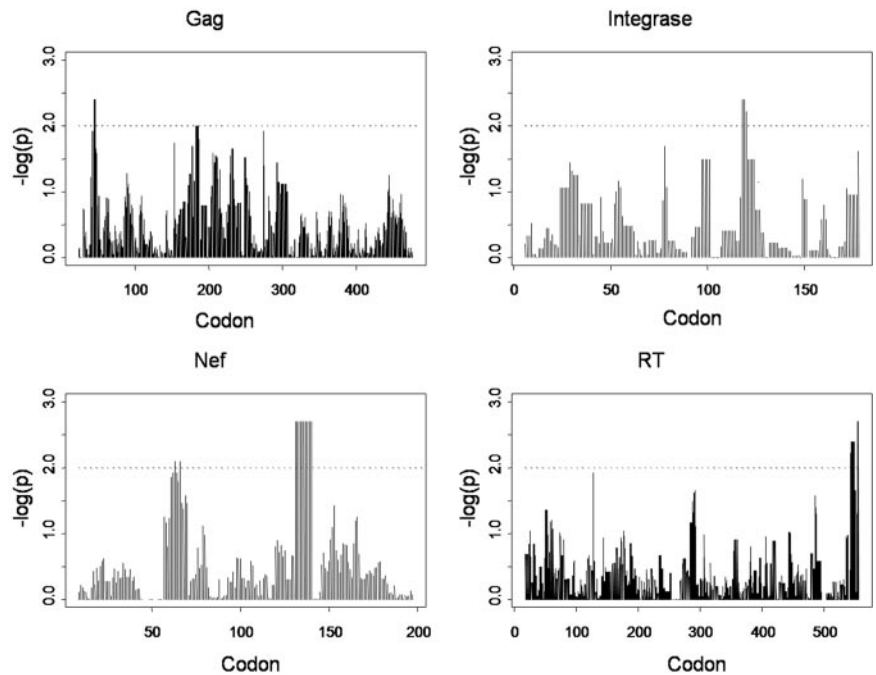
viral diversity in the population is strongly determined by CTL selection. In contrast, the divergences over whole proteins or longer lengths of viral sequence encompass multiple immune and nonimmune influences (including lineage).

Cumulative HLA-driven adaptation per person

To determine the extent of cumulative adaptation on a per-host basis, we examined, in each person, the viral residues in their autologous sequence at which a significant HLA-HIV polymorphism association had been detected in the previously described population-level analysis. Because the number of potential adaptation sites will probably be determined by a person's HLA genotype, we calculated the number of residues in the HLA-adapted state as a proportion of total residues potentially associated with adaptation to that person's own HLA-A, -B, and

-C alleles (Figure 4). Therefore, persons with no relevant HLA-association sites by virtue of their particular HLA genotypes were excluded. Although the extremes of association were based on small numbers, in the middle ranges containing more persons there was a strikingly constant percentage of cumulative adaptation, between 30% and 60%, for Gag, RT, integrase, and Nef, regardless of the number of sites potentially subject to adaptation. For example, in Gag, an average of 50% of sites subject to HLA-driven change were adapted in persons, whether those persons had only 5 or > 15 residues potentially subject to HLA-associated pressure. Notably, there was significant cumulative adaptation occurring in persons with many available sites over all proteins, suggesting that CD8 T cells exert selective pressure across multiple epitopes within persons with chronic HIV-1 infection.
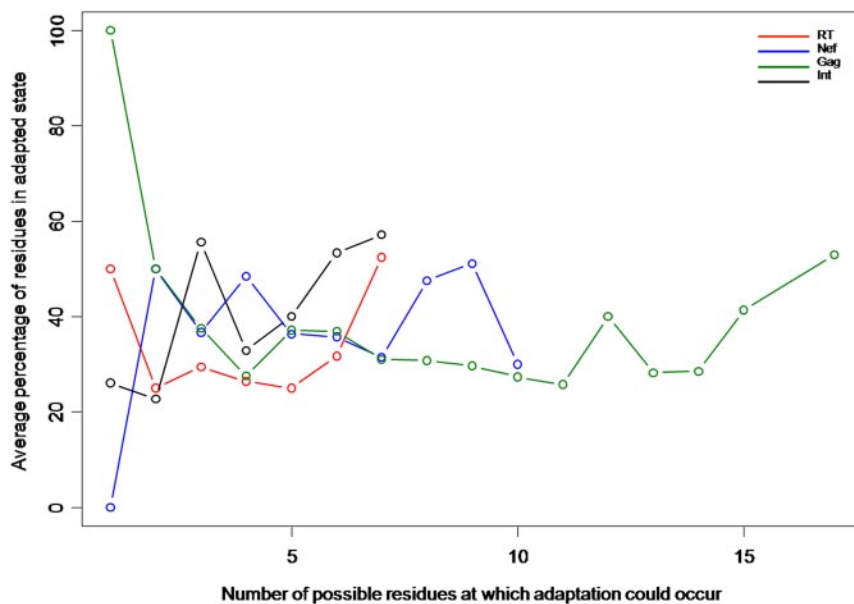

Figure 4. Plot of mean percentage cumulative HLA-associated adaptation per sequence (y-axis) in persons in the SM cohort according to the number of residues potentially subject to HLA-associated adaptation (x-axis). Individual plots for each protein are examined.
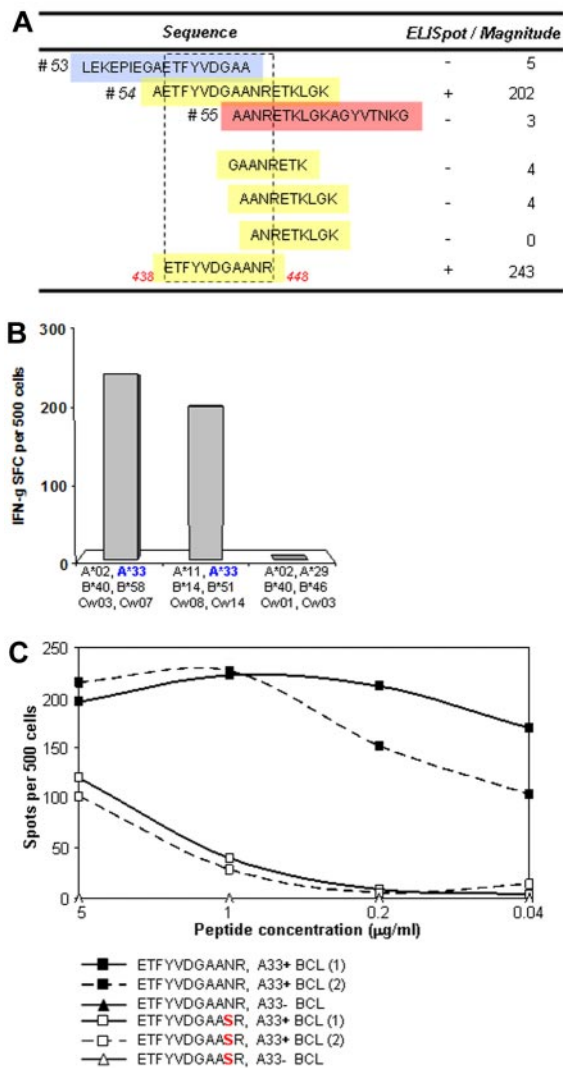
Figure 5. T-cell assays to determine whether a strongly A33-associated mutation in RT lies in a novel epitope. (A) T-cell lines from A33+ donors were established by stimulation with an 18-mer peptide reflecting the consensus sequence (pol 54) and tested for recognition against A33-matched targets pulsed with the overlapping peptides (pol 53 and 55, top) and with truncated peptides (bottom) to define the optimal epitope. (B) Confirmation of HLA-A33 restriction was performed with target cell lines matched only at HLA-33 or lacking HLA-A33. (C) Confirmation that the N447S mutation represents an escape from T-cell recognition was performed with pol 54 peptide-specific T-cell clones tested for recognition of the wild-type and mutant peptides in an ELISPOT assay with 2 different A33-expressing target cells.

HLA-associated mutations predict the presence of previously unknown CTL epitopes

A substantial proportion of HLA-associated mutations did not lie close to or within known CTL epitopes. However, because only a limited amount of CTL epitope mapping has been performed in Chinese cohorts for HLA alleles common in the Chinese population,[35,36] these mutations could be markers for previously unidentified epitopes. For an HLA-A33–associated mutation in Pol, we identified donors with HLA-A33 and responses to the consensus (ie, nonadapted) 18-mer sequence containing the residue of interest, in whom the viral sequence did not show a mutation at this time point. CTL lines and clones were established with the 18-mer peptide and used to determine the optimal epitope and restricting HLA molecule. These studies showed the presence of a previously unknown HLA-A33–restricted epitope in RT (Figures 5A-B). The A33-associated mutation, RT N447S, was detected in 10 of

22 subjects with HLA-A33 but in only 3 of 69 donors without the A33 allele. None of the A33 donors with the RT N447S mutation made a T-cell response to the Pol54 consensus peptide, nor did CTL responding to the nonmutated epitope recognize the variant peptide (Figure 5C), confirming that this sequence change represents a CTL-driven escape variant.

Discussion

In this epidemiologically unique population, sequences across 3 major HIV-1 genes formed a monophyletic subtype B′ cluster that cosegregated with other sequences associated with the plasma donation epidemic in central China.[22] This pattern suggests it is unlikely that ingress from the sexual transmission populations or injecting drug user populations elsewhere in China had occurred in this study cohort. The sequences exhibited relatively restricted genetic diversity, comparable to intrahost divergence over time, and approximately one-half that seen in a typical population-based cohort in the United States with a longer history of HIV-1 infection, presumably with multiple founders and complex transmission networks. The age of the most recent common ancestor for this cluster coincides with the relatively short interval during which the plasma donation clinics operated in this particular village, arguing that this outbreak represents a more focused geographic sampling within the wider monophyletic plasma donation epidemic involving several cities and villages. Taken together these results suggest to a high degree of certainty that this outbreak arose from a narrow source, in keeping with all the available epidemiologic information on HIV-1 infection within this population. Analysis of interactions between HLA and HIV show that pairwise divergence in HLA genotype correlates with pairwise divergence in viral sequence, although this correlation is, as might be predicted, only evident within localized windows that reflect the epitope targeting of HLA class I–restricted CTL. We show that many sites of viral polymorphism within these windows are HLA allele specific and that the most intense selection effects are associated with HLA-B locus alleles. Of the 3 HIV-1 proteins studied, all of which elicit potent CTL responses, selection was most apparent for Nef, consistent with previous studies,[5,18,19,32] presumably because the functional and structural constraints on mutation are greater for Gag and Pol. We estimate that HLA-associated mutations account for between 24% and 56% of the polymorphic sites detected in gag, nef, and pol sequences in this cohort, showing the extensive contribution of cellular immune selection to viral evolution. At the per-host level, all proteins show extensive HLA-associated adaptation in the order of 30%-60% of sites subject to CTL selection. The biologic significance of HLA associations is supported by the demonstration that HLA-associated mutations lie in previously undefined T-cell epitopes restricted by these HLA molecules, as shown here in the example of a novel HLA-A33–restricted Pol epitope.

Outbreaks of monophyletic HIV-1 strains are rare but have been described previously in injecting drug users in Kaliningrad[37] and in children attending a Libyan hospital.[38] However, there are additional features of the SM cohort that distinguish it from other narrow source epidemics. It is probable that cohort members were infected within a relatively short time frame, because all cohort members were plasma donors and few cases of HIV-1 infection have been detected in nondonors in the village. Because HIV-1 infection was not diagnosed and treated until 2004, this permits the direct observation of viral diversification from a narrow source over almost 10 years without the confounding influence of ART.

The cohort is ethnically homogeneous Han Chinese, so the HLA repertoire influencing viral evolution is well defined and relatively limited. Inevitably the cohort is restricted to those who survived until 2004, thereby limiting the information that can be gleaned about rapid progression in the villagers who died before this date. Nevertheless, a cohort in which several of the major variables that affect the natural history of HIV-1 infection (viral strain, route and timing of infection, and ethnic diversity) are controlled provides an unparalleled opportunity to determine host genetic factors that influence clinical outcome, which will be the basis of future studies.

These data confirm the central role of CD8 T cells restricted by class I HLA molecules in driving viral evolution. Since the first description of the emergence of viral variants that escape T-cell recognition in chronic HIV-1 infection,[6] evidence has accumulated that selection by HIV-specific CTLs contributes to viral variation, but it has been difficult to quantify this contribution accurately. It is now clear that CTLs drive viral diversification from early stages of infection,[16] but escape may also occur late in infection, when it has been associated with clinical deterioration.[7] Late escape may be a consequence of the number and complexity of the mutations required to generate a replication-competent T-cell escape variant, as in the case of the immunodominant HLA-B27–restricted epitope in Gag, KK10, for which a combination of 3 amino acid substitutions are required (one of which is outside the epitope)[39]; an additional compensatory mutation restores replication capacity close to the wild-type level.[40] The long-term stability of T-cell escape mutants depends on the fitness cost incurred by the virus; variants with a high-fitness cost tend to revert to the original sequence after transmission to a host without the selecting HLA allele, unless an appropriate compensatory mutation is also present. Other variants revert only slowly, if at all,[41] thereby potentially compromising the efficacy of the CTL response to HIV-1 in donors with the selecting HLA types. At a population level, the accumulation of escape mutations for a HLA-B51–restricted response reflects the prevalence of this allele in the population; in Japan, where HLA-B51 is most common, this has been sufficient to undermine the earlier association of HLA-B51 with viral control.[42] Selection of variants with a high-fitness cost in conserved regions of the virus has been proposed as an important mechanism to explain the association of certain HLA class I molecules such as HLA-B57 and -B27 with delayed disease progression in HIV-1 infection.[43] Examination of HIV-1 transmission pairs suggests that primary infection with CTL escape variants is advantageous to HLA-mismatched recipients,[44] presumably because the transmitted virus has some degree of impaired replicative capacity. The implications of HLA-mediated viral selection are therefore complex; although in some cases the adapted virus may have lost susceptibility to the host's most potent antiviral CTL, reduced replicative capacity of some escape variants may provide a relative advantage to the host. Moreover, because dominant T-cell responses are lost at the population level because of viral adaptation, the development of subdominant responses may confer enhanced viral control, as has been shown in other viral infections.[45]

The stability, ethnic homogeneity, and high degree of interrelatedness of the SM village population results in viral selection being subject to a relatively limited number of HLA alleles; however, this leads to the associations between viral polymorphisms and individual HLA alleles present at high frequency being very strong, such as the HLA-A24 association with characteristic mutations at position 133 and 135 in Nef, which lie within or close to an immunodominant HLA-A24–restricted nef epitope.[46] We have shown that strong class I HLA associations that are not in known CTL epitopes may predict novel epitopes, which is particularly valuable for less-studied vaccine target populations. Nevertheless, there is a substantial amount of individual viral variation that is not explained by HLA-mediated effects. Future studies may identify other immune influences on viral evolution; for example, natural killer cells are able to respond to individual HIV peptides,[47] and interactions between killer immunoglobulin-like receptor molecules expressed on natural killer cells and HLA class I molecules are sensitive to the peptides bound to the HLA molecule, including HIV epitope peptides.[48]

The rapid adaptation of HIV-1 to evade CD8+ T-cell responses at both the individual and population levels has significant implications for vaccine strategies. Although recent studies of T cell–inducing vaccines have shown encouraging results in macaque models,[49,50] it remains a major challenge to generate a T-cell response in humans that will provide protection against diverse strains of HIV-1, especially when HIV-1 strains circulating within a population have acquired stable mutations to evade the dominant T-cell responses restricted by common HLA molecules in that population. However, combining the analytical approaches presented here with T-cell studies across the viral proteome in populations such as the SM cohort should allow the determination of regions of the virus that are both immunogenic and subject to functional or structural constraints or both, which are the most likely to elicit potentially protective immune responses.

Acknowledgments

Authorship

Contribution: T.D., S.L.R.-J., Y.H.Z., and K.Y.X. designed the study; Y.H.Z., H.P.Y., Y.C.P., M.-E.B., T.D., H.W., X.Y.C., Y.M., N.L., W.Y.Q., W.H.L., T.R., and X.X. performed the experiments and were involved in patient recruitment; T.D., S.L.R.-J., Y.H.Z., M.J., I.J., and S.G. performed data analysis; and S.L.R.-J., M.J., T.D., A.M., and S.M. wrote the paper.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Sarah L. Rowland-Jones, MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, Oxford, OX3 9DS United Kingdom; e-mail: sarah.rowland-jones@ndm.ox.ac.uk; and Tao Dong, MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, Oxford, OX3 9DS United Kingdom; e-mail: tao.dong@imm.ox.ac.uk.

References

1. McMichael AJ, Rowland-Jones SL. Cellular immune responses to HIV. *Nature.* 2001;410(6831): 980-987.

2. McElrath MJ, De Rosa SC, Moodie Z, et al. HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis. *Lancet.* 2008; 372(9653):1894-1905.

3. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol.* 1999;73(12):10489-10502.

4. Malim MH, Emerman M. HIV-1 sequence variation: drift, shift, and attenuation. *Cell.* 2001;104(4): 469-472.

5. Brumme ZL, John M, Carlson JM, et al. HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One.* 2009;4(8):e6687.

6. Phillips RE, Rowland-Jones SL, Nixon DF, et al. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature.* 1991;354(6353):453-459.

7. Goulder PJ, Phillips RE, Colbert RA, et al. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat Med.* 1997;3(2):212-217.

8. Leslie AJ, Pfafferott KJ, Chetty P, et al. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med.* 2004;10(3):282-289.

9. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science.* 2002;296(5572):1439-1443.

10. Leslie A, Kavanagh D, Honeyborne I, et al. Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J Exp Med.* 2005;201(6): 891-902.

11. Kawashima Y, Pfafferott K, Frater J, et al. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature.* 2009;458(7238):641-645.

12. Bhattacharya T, Daniels M, Heckerman D, et al. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science.* 2007;315(5818):1583-1586.

13. Scherer A, Frater J, Oxenius A, et al. Quantifiable cytotoxic T lymphocyte responses and HLA-related risk of progression to AIDS. *Proc Natl Acad Sci U S A.* 2004;101(33):12266-12270.

14. Asquith B, McLean AR. In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proc Natl Acad Sci U S A.* 2007;104(15):6365-6370.

15. Borrow P, Lewicki H, Wei X, et al. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat Med.* 1997;3(2):205-211.

16. Goonetilleke N, Liu MK, Salazar-Gonzalez JF, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med.* 2009; 206(6):1253-1272.

17. Price DA, Goulder PJ, Klenerman P, et al. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc Natl Acad Sci U S A.* 1997;94(5):1890-1895.

18. Rousseau CM, Daniels MG, Carlson JM, et al. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *J Virol.* 2008;82(13): 6434-6446.

19. Brumme ZL, Brumme CJ, Heckerman D, et al. Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathog.* 2007;3(7):e94.

20. Matthews PC, Leslie AJ, Katzourakis A, et al. HLA footprints on human immunodeficiency virus type 1 are associated with interclade polymorphisms and intraclade phylogenetic clustering. *J Virol.* 2009;83(9):4605-4615.

21. Kaufman J, Jing J. China and AIDS–the time to act is now. *Science.* 2002;296(5577):2339-2340.

22. Zhang L, Chen Z, Cao Y, et al. Molecular characterization of human immunodeficiency virus type 1 and hepatitis C virus in paid blood donors and injection drug users in china. *J Virol.* 2004;78(24): 13591-13599.

23. Dong T, Stewart-Jones G, Chen N, et al. HIV-specific cytotoxic T cells from long-term survivors select a unique T cell receptor. *J Exp Med.* 2004; 200(12):1547-1557.

24. Scheider S, Kueffer JM, Roessli D, Excoffier L. Arlequin: a software for population genetic data analysis, version 1.1. Genetics and Biometry Laboratory Department of Anthropology, University of Geneva. 1997.

25. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 2001;125(1-2):279-284.

26. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* 2003;100(16):9440-9445.

27. Dorrell L, Willcox BE, Jones EY, et al. Cytotoxic T lymphocytes recognize structurally diverse, clade-specific and cross-reactive peptides in human immunodeficiency virus type-1 gag through HLA-B53. *Eur J Immunol.* 2001;31(6):1747-1756.

28. Li S, Jiao H, Yu X, et al. Human leukocyte antigen class I and class II allele frequencies and HIV-1 infection associations in a Chinese cohort. *J Acquir Immune Defic Syndr.* 2007;44(2):121-131.

29. Hong W, Fu Y, Chen S, Wang F, Ren X, Xu A. Distributions of HLA class I alleles and haplotypes in Northern Han Chinese. *Tissue Antigens.* 2005;66(4):297-304.

30. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics.* 1992;48(2):361-372.

31. Su B, Liu L, Wang F, et al. HIV-1 subtype B′ dictates the AIDS epidemic among paid blood donors in the Henan and Hubei provinces of China. *AIDS.* 2003;17(17):2515-2520.

32. John M, Heckerman D, James I, et al. Adaptive interactions between HLA and HIV-1: highly divergent selection imposed by HLA class I molecules with common supertype motifs. *J Immunol.* 2010; 184(8):4368-4377.

33. Hue S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A.* 2005;102(12):4425-4429.

34. Rauch A, James I, Pfafferott K, et al. Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen-restricted immune pressure. *Hepatology.* 2009;50(4):1017-1029.

35. Gong X, Gui X, Zhang Y, Tien P. Screening for CD8 cytotoxic T lymphocytes specific for Gag of human immunodeficiency virus type 1 subtype B′ Henan isolate from China and identification of novel epitopes restricted by the HLA-A2 and HLA-A11 alleles. *J Gen Virol.* 2006;87(Pt 1):151-158.

36. Zhai S, Zhuang Y, Song Y, et al. HIV-1-specific cytotoxic T lymphocyte (CTL) responses against immunodominant optimal epitopes slow the progression of AIDS in China. *Curr HIV Res.* 2008; 6(4):335-350.

37. Liitsola K, Tashkinova I, Laukkanen T, et al. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS.* 1998;12(14):1907-1919.

38. de Oliveira T, Pybus OG, Rambaut A, et al. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature.* 2006;444(7121): 836-837.

39. Kelleher AD, Long C, Holmes EC, et al. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *J Exp Med.* 2001;193(3): 375-386.

40. Schneidewind A, Brockman MA, Yang R, et al. Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J Virol.* 2007;81(22):12382-12393.

41. Kearney M, Maldarelli F, Shao W, et al. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J Virol.* 2009;83(6):2715-2727.

42. Kawashima Y, Kuse N, Gatanaga H, et al. Long-term control of HIV-1 in hemophiliacs carrying slow-progressing allele HLA-B*5101. *J. Virol.* 2010;84(14):7151-7160.

43. Martinez-Picado J, Prado JG, Fry EE, et al. Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *J Virol.* 2006;80(7):3617-3623.

44. Chopera DR, Woodman Z, Mlisana K, et al. Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. *PLoS Pathog.* 2008;4(3):e1000033.

45. Holtappels R, Simon CO, Munks MW, et al. Subdominant CD8 T-cell epitopes account for protection against cytomegalovirus independent of immunodomination. *J Virol.* 2008;82(12):5781-5796.

46. Ikeda-Moore Y, Tomiyama H, Miwa K, et al. Identification and characterisation of multiple HLA-A24-restricted HIV-1 CTL epitopes: strong epitopes are derived from V regions of HIV-1. *J Immunol.* 1997;159:6242-6252.

47. Stratov I, Chung A, Kent SJ. Robust NK cell-mediated human immunodeficiency virus (HIV)-specific antibody-dependent responses in HIV-infected subjects. *J Virol.* 2008;82(11):5450-5459.

48. Thananchai H, Gillespie G, Martin MP, et al. Cutting Edge: allele-specific and peptide-dependent interactions between KIR3DL1 and HLA-A and HLA-B. *J Immunol.* 2007;178(1):33-37.

49. Liu J, O'Brien KL, Lynch DM, et al. Immune control of an SIV challenge by a T-cell-based vaccine in rhesus monkeys. *Nature.* 2009;457(7225):87-91.

50. Hansen SG, Vieville C, Whizin N, et al. Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nat Med.* 2009;15(3):293-299.