Coordinate linkage of HIV evolution reveals regions of immunological vulnerability

Vincent Dahirel[a,b,c,1], Karthik Shekhar[a,b,1], Florencia Pereyra[a], Toshiyuki Miura[d], Mikita Artyomov[c,e], Shiv Talsania[b,f], Todd M. Allen[a], Marcus Altfeld[a], Mary Carrington[a,g], Darrell J. Irvine[a,h,i], Bruce D. Walker[a,h,2] and Arup K. Chakraborty[a,b,c,i,2]

[a]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, MA 02129; Departments of [b]Chemical Engineering, [c]Chemistry, and [i]Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Department of Chemistry, Moscow State University, Moscow 119991, Russia; [f]Department of Chemical Engineering, Loughborough University, Leicestershire LE11 3TU, United Kingdom; [g]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, MD 21702; [h]Howard Hughes Medical Institute, Chevy Chase, MD 20815; and [d]Institute for Medical Sciences, University of Tokyo, Tokyo 108-8639, Japan

Edited* by Laurie H. Glimcher, Harvard University, Boston, MA, and approved May 20, 2011 (received for review April 4, 2011)

Cellular immune control of HIV is mediated, in part, by induction of single amino acid mutations that reduce viral fitness, but compensatory mutations limit this effect. Here, we sought to determine if higher order constraints on viral evolution exist, because some coordinately linked combinations of mutations may hurt viability. Immune targeting of multiple sites in such a multidimensionally conserved region might render the virus particularly vulnerable, because viable escape pathways would be greatly restricted. We analyzed available HIV sequences using a method from physics to reveal distinct groups of amino acids whose mutations are collectively coordinated ("HIV sectors"). From the standpoint of mutations at individual sites, one such group in Gag is as conserved as other collectively coevolving groups of sites in Gag. However, it exhibits higher order conservation indicating constraints on the viability of viral strains with multiple mutations. Mapping amino acids from this group onto protein structures shows that combined mutations likely destabilize multiprotein structural interactions critical for viral function. Persons who durably control HIV without medications preferentially target the sector in Gag predicted to be most vulnerable. By sequencing circulating viruses from these individuals, we find that individual mutations occur with similar frequency in this sector as in other targeted Gag sectors. However, multiple mutations within this sector are very rare, indicating previously unrecognized multidimensional constraints on HIV evolution. Targeting such regions with higher order evolutionary constraints provides a novel approach to immunogen design for a vaccine against HIV and other rapidly mutating viruses.

cytotoxic T-lymphocyte response | elite controllers | random matrix theory

Despite the efficacy of life-extending medications, HIV continues to wreak havoc around the world, particularly in poor nations. An efficient vaccine is urgently needed, and such a vaccine is likely to be one that induces both antibody and cytotoxic T-lymphocyte (CTL) responses (1–3). The extreme variability of the HIV envelope has precluded vaccine-induced generation of broadly neutralizing antibodies (1) that can prevent acquisition, leading some to focus on CTL-based vaccines (4) to prevent disease progression and possibly to limit acquisition (5).

CTLs recognize and respond to viral peptides presented by class I MHC proteins. Single point mutations within and surrounding such HIV epitopes targeted by CTLs can enable escape from immune pressure, leading to a focus on targeting conserved regions of the HIV proteome. Conserved regions are defined to be ones where the frequency of occurrence of mutations at single sites is small, indicating [according to evolutionary theory (6)] that the corresponding mutant viral strains are replicatively less fit. Thus, if such a site is targeted, the outgrowth of a mutant virus that escapes the immune pressure is less likely (7). The emergence of a compensatory mutation that restores fitness is a challenge to this approach (8).

Characterizing the frequency of occurrence of viral strains based on the effects of mutations at single sites results in a unidimensional measure of conservation, which ignores potential couplings between the effects of multiple simultaneous mutations due to structural/functional constraints. A useful multidimensional measure of sequence variation would be obtained if groups of sites in the viral proteome could be identified, such that sites within a group coevolve in a collectively interdependent manner to influence virus viability but each group evolves independent of other groups. If such groups exist, it is possible that a greater proportion of the combinations of mutations involving sites in the group are harmful in some groups compared with other groups. Such a group of sites is multidimensionally more conserved in that multiple mutations are more likely to hurt virus fitness and harmful combinations are less likely to be compensated for by mutations in another group (as it coevolves independently). Such regions should be particularly vulnerable to multiple points of CTL pressure, because escape pathways would be restricted. Targeting multiple points would promote the emergence of multiple mutations to escape the immune pressure, but multiple mutations in such a region would be more likely to result in unfit viruses.

Thus, we sought to determine collectively coevolving groups of residues in HIV proteins. We analyzed publicly available sequences from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/) of clade B HIV proteins derived from patients with diverse genotypes using random matrix theory (RMT) (9) to identify groups of sites (not pairs) that evolve collectively, presumably because they act together to perform a function important for the virus. RMT has been applied to diverse realms of physics, to analyze stock price fluctuations (10, 11), and to analyze sequences of an enzyme (12). Although inspired by Halabi et al. (12), we outline our analysis of HIV proteins by analogy with analyses of financial markets.

From price fluctuations of various stocks over time, one can compute the extent of correlation between changes in price of a pair of stocks, averaged over all available time points. This yields a matrix of pair correlations. Each element of the matrix describes correlations between a particular pair of stocks, and mathematical

Author contributions: V.D., K.S., and A.K.C. designed research; V.D., K.S., and A.K.C. performed research; F.P., T.M., M.C., and B.D.W. contributed new reagents/analytic tools; V.D., K.S., M. Artyomov, S.T., B.D.W., and A.K.C. analyzed data; and V.D., K.S., T.M.A., M. Altfeld, D.J.I., B.D.W., and A.K.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]V.D. and K.S. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: bwalker@partners.org or arupc@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1105315108/-/DCSupplemental.

properties of the entire matrix (*Results*) reveal groups of stocks whose price fluctuations are collectively interdependent. These properties of the matrix are influenced by "noise" because of a finite sample of times and correlations that reflect changes in overall stock prices attributable to external forces (e.g., recessions). RMT has been used to "clean" the correlation matrix of these effects and obtain groups of companies whose economic activities are intrinsically collectively coupled and essentially independent of others (10, 11). Consistent with intuition, each group was composed of stocks that define known economic "sectors," such as financial service companies and automotive companies.

A similar analysis has enabled us to identify collectively coevolving groups of sites in the HIV proteome (which we term HIV sectors). Further, we have identified a sector in which mutations at multiple sites are collectively more constrained, and thus might be particularly vulnerable to multiple points of immune pressure. We examined immune targeting and sequenced circulating viruses in those rare persons who are able to control HIV without medications to determine whether such mechanisms might contribute to durable immune control. These data support our predictions. We suggest how this new understanding can be harnessed to design immunogens for a vaccine against HIV.

Results

Sequence Analysis Identifies Significant HIV Sectors. Each aligned sequence of an HIV polyprotein is analogous to stock price data at a particular time point. In a specific sequence, one asks if a mutation away from the most frequent amino acid ("wild type") at a particular position i appears simultaneously with a similar mutation at another site j, and this is done for all pairs of amino acids. Average values for the frequency of occurrence of double mutations for each pair of sites in the protein (f_{ij} for sites i and j) are obtained by repeating this procedure for all sequences (*Methods* and *SI Appendix* 1). Similarly, the frequencies with which mutations are observed at individual sites (f_i and f_j for sites i and j) are obtained. A pair correlation matrix, **C**, can be

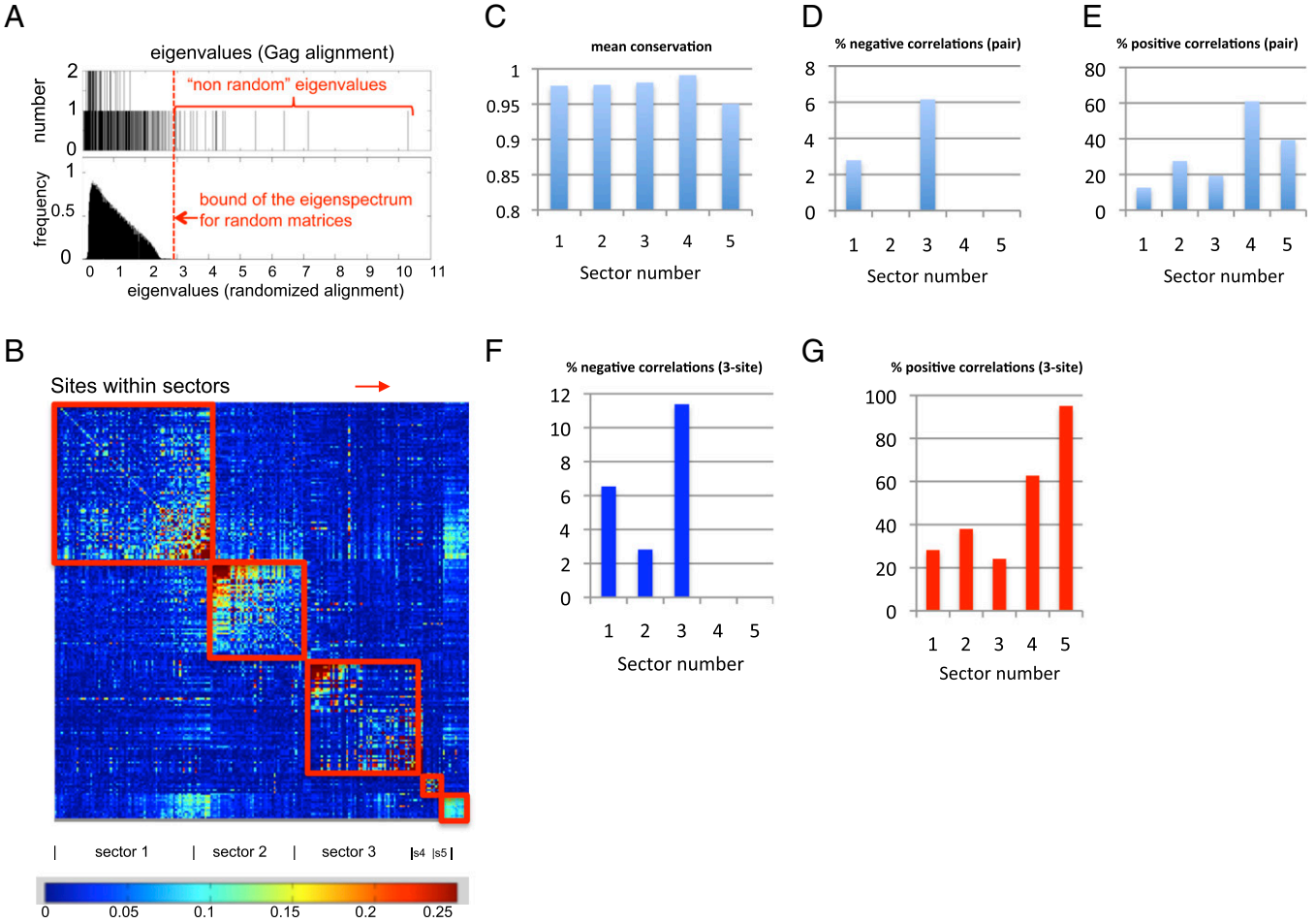

Fig. 1. Defining collectively coevolving groups of sites (sectors) in Gag. (*A, Upper*) Number of eigenvalues (ordinate) of the correlation matrix, **C**, for Gag (defined in the main text) with a given magnitude (abscissa) is shown. In total, 1,600 sequences of Gag polyproteins (500 residues long) were used. Note that a relatively large number of sequences of HIV proteins are available. (*Lower*) Distribution of eigenvalues obtained from 1,000 randomly generated matrices of the same size as **C** for Gag (described in the main text and *SI Appendix* 2). (*B*) Analyzing the eigenvectors corresponding to eigenvalues larger than the highest eigenvalue for random matrices yields collectively coevolving groups of sites or "sectors" (*SI Appendix* 3–7). Sites within a sector are grouped together along the rows and columns so that groups of collectively coevolving sites are vivid as squares along the diagonal of a heat map representing the values of the correlations (obtained from the "cleaned" correlation matrix as described in *SI Appendix* 2). (*C*) Mean frequency of the dominant amino acid at single sites within each of the five Gag sectors. (*D–F*) Threshold value defining a significant correlation was chosen to be such that correlations with magnitudes greater than this value arise with vanishing probability in the randomized matrices ($P < 0.02$). Changing this threshold value does not change qualitative results (*SI Appendix*, Fig. S11). (*D*) Percentage of significant negative ($C_{ij} < -0.03$) pair correlations within each sector. (*E*) Percentage of significant positive ($C_{ij} > 0.1$) pair correlations within each sector. (*F*) Percentage of significant three-site negative ($C_{ijk} < -0.01$) correlations within each sector (method in *SI Appendix* 17). (*G*) Percentage of significant three-site positive ($C_{ijk} > 0.1$) correlations within each sector.

defined, each element of which, C_{ij} (for sites i and j), reflects the interdependence of mutations at these two sites. $C_{ij} = (f_{ij} - f_i f_j)/\sqrt{V_i V_j}$ and measures the difference between the frequency of occurrence of a double mutant and that which would be observed if the two mutations occurred independently; the variances of the distributions of mutations (V_i and V_j) at sites i and j, respectively, normalize the values of C_{ij} for different pairs of sites. Another method (12) for computing C_{ij} yielded qualitatively similar results (*SI Appendix*, Fig. S6).

Properties of this correlation matrix, termed eigenvectors, contain information about collective coevolution of sites. Each eigenvector represents a specific combination of sites, whose mutations occur in a coupled way but are essentially independent of mutations in a combination of sites represented by another eigenvector. The data on HIV evolution can be represented as dependent on these combinations of sites (eigenvectors), rather than individual sites. The association of each site with a particular eigenvector is specified by a number that can be positive or negative. An eigenvector is also associated with a number, called an eigenvalue, whose magnitude reflects the contribution of this eigenvector to the correlations between sites (*SI Appendix*, Eq. **S4**).

Eigenvalues of the correlation matrix, C, for HIV Gag polyproteins are shown ordered according to their magnitudes (Fig. 1*A*, *Upper*). The corresponding eigenvectors describe groups of sites whose mutations are collectively linked, presumably by the requirements of maintaining viral fitness. However, the information in some eigenvectors is not significant because of noise (attributable to a finite sample of sequences) or because it reflects phylogeny (13).

RMT theorems can help to determine which eigenvectors reflect the influence of noise because they describe the properties of correlation matrices derived from independent random variables. For example, given the length of the Gag polyprotein and the number of available sequences, RMT states (*SI Appendix*, Eq. **S3**) that the largest eigenvalue would equal 2.4 if the correlation matrix reflected noise only. However, this theorem assumes that the length of the polyprotein and the number of sequences are very large. If we randomly shuffle the identity of amino acids at each site over all aligned sequences of a particular HIV protein and recalculate C, we obtain a random correlation matrix of the same size as our sample (*SI Appendix* 2). The distribution of eigenvalues obtained by randomizing the sequences 1,000 times is shown in Fig. 1*A*, *Lower*. We see that the RMT theorem is rather accurate even for finite samples, because very few eigenvalues exceed 2.4 and none exceed 3. The continuous part of the eigenvalue spectrum for the real-sequence alignment for HIV proteins is bounded in the same way as that for the randomized sequences (compare panels in Fig. 1*A* for Gag). The eigenvectors corresponding to eigenvalues <3 correspond to noise.

As noted before (12), and derived more completely in *SI Appendix* 3, the contribution of phylogeny alone to a representation of the data in terms of eigenvectors should result in an eigenvector where the contributions from each individual site in the polyprotein have the same sign. For all HIV polyproteins, the largest eigenvalue corresponds to such an eigenvector. Because each eigenvector is independent, we conclude that the eigenvector corresponding to the largest eigenvalue describes correlations attributable to phylogeny (13). Eigenvectors corresponding to the next few largest eigenvalues contain information on collective correlations between groups of sites that originate largely from relationships between evolutionary constraints and sequence.

By parsing these eigenvectors, groups of sites (sectors) in HIV proteins that coevolve together but essentially independent of others were identified. Many sites do not collectively coevolve with any other sites, and thus do not belong to any sector. For Gag, we found five strongly collectively coupled groups of sites (*SI Appendix* 6 and Fig. S4), which we term sectors 1–5. As expected, we also found an additional group of residues that is weakly coupled collectively, as a result of mutations arising in the context of specific HLA class I molecules that drive mutations at multiple sites (*SI Appendix* 8 and Figs. S4 and S5). The weakly collective linkage between this group of sites is attributable to HLA-associated "footprints" [i.e., these mutations arise in persons with the same HLA type (14)]; thus, we do not consider this quasi-sector further.

Gag sectors 1–5 can be visualized as a heat map reflecting the correlations between sites (Fig. 1*B*). As expected, the sites that comprise a sector are not contiguous along the linear protein sequence. The positions of the sites along the rows and columns in Fig. 1*B* group the sites in a sector together. For example, if a sector was composed of sites 27, 45, and 53 along the linear protein sequence, these sites would be adjacent to each other along a row or column. Each group of sites that comprises a sector is outlined in red. We observed similar groupings for other HIV proteins (*SI Appendix*, Figs. S7–S10). We focus further analysis on Gag because of data indicating correlations between Gag-specific responses and viral control (15) and a strong impact of Gag mutations on viral fitness (16).

Identification of an Immunologically Vulnerable Sector in Gag. From the standpoint of identifying regions with single sites that are more conserved, all five Gag sectors are equivalent (Fig. 1*C*). However, if one examines the relative occurrence of positive and negative correlations between multiple mutations at sites that comprise each collectively evolving sector, differences between the sectors become apparent.

Negative correlation between sites implies that the multiple mutant is observed less frequently than if the individual mutations were to arise independently. For example, consider a simple case where a pair of sites (i and j) is negatively correlated, with each site being 90% conserved [i.e., the frequency with which single mutations at sites i and j are observed is 10%, ($f_i = f_j = 0.1$)]. For C_{ij} (proportional to $f_{ij} - f_i f_j$) to be negative, the frequency with which the double mutant is observed (f_{ij}) must be less than 1%. This implies that when multiple sites are negatively correlated, the multiple mutant is considerably less fit compared with the single mutants. If immune pressure is applied to multiple sites in a sector, escaping the immune pressure is likely to require more than one mutation. The greater the proportion of negative correlations in the sector, the more difficult it is to both escape the CTL pressure and maintain virus fitness, because multiple mutations are much less tolerated, likely because of fitness limitations.

Positive correlation between sites implies that the corresponding multiple mutants are observed more frequently than if the mutations were to occur independently. Thus, positive correlations can be associated with compensatory mutations; indeed, known compensatory mutations do appear as positive correlations (*SI Appendix*, Table S9). Thus, sectors characterized by larger numbers of positive correlations would be less effective targets, because escape from CTLs without substantial loss in virus fitness is more likely.

For pairs of sites within a sector, one finds that Gag sector 3 contains the largest proportion of negative correlations compared with other sectors and a relatively small number of positive correlations (Fig. 1 *D* and *E*). This is also true for three-site (Fig. 1 *F* and *G*) and higher order correlations. The magnitudes of the negative pair correlations between sites in sector 3 are as large as they can be, given the level of single site conservation in this sector (details in *SI Appendix* 11), meaning that the double mutants are expected to be observed very rarely. The eigenmaps in *SI Appendix*, Fig. S4 *B* and *C* also show that several sites in sector 3 are collectively (i.e., at higher order) negatively correlated to several others, which is not true for other sectors. Note that we do not identify important negative pair correlations by screening all pair correlations for those that exceed a cutoff (17). Rather, we first identified groups of sites (sectors) that are significantly collectively coupled (because RMT theorems help to eliminate noise-induced

correlations). We then examined the signs of multibody correlations only within these meaningful collectively coupled sectors.

Our results suggest that sector 3 is the most immunologically vulnerable multidimensionally constrained region in Gag, because multiple mutations are most constrained in this sector. Because sequence analysis methods are not exact, we tested this prediction, and its consequences for HIV infection and vaccination, against existing and new experimental data.

Sector 3 Is Immunologically Vulnerable Because of the Importance of Assembling Multiprotein Structures Critical for Viral Capsid Formation. In sector 3, 52 of the 57 sites are contained in the p24 protein but their locations within an individual p24 protein appear random (Fig. 2*A*; individual amino acids in *SI Appendix*, Table S1). However, hexamers of the p24 protein form the viral capsid (18), and superimposing sector 3 sites on the structure of the p24 hexamer (19) shows that the preponderance of these sites (~68%) is at interfaces between p24 proteins in a hexamer or at interfaces between the hexamers that form the capsid (Fig. 2*B*). Coevolution of this group of sites originates from constraints important for assembly of functional multiprotein structures, highlighting the importance of determining collective correlations.

The structural locations of sector 3 sites suggest the origin of negative correlations. Whereas one mutation in interface residues may still allow formation of p24 hexamers and assembly of the hexamers to form the viral capsid, multiple simultaneous mutations would likely destabilize these protein-protein interfaces. Fitness cost predictions of multiple mutations in sector 3, and comparisons with available data (20), are included in *SI Appendix*, Table S9–S11.

Sector 1 sites also reflect structural constraints associated with supramolecular assembly because they largely comprise the core of the hexamer (Fig. 2*C*). However, our analyses suggest that sector 1 is not as immunologically vulnerable as sector 3. This is because of the following:

i) The ratio of negative to positive correlations for pairs of mutations is a factor of 3 greater for sector 3.
ii) At the level of three-site correlations, the ratio of negative to positive correlations is more than threefold greater for sector 3.
iii) Sector 3 is characterized by collective higher order negative correlations between sites that are absent in sector 1 (*SI Appendix*, Fig. S4).

These results may be because multiple mutations in sector 3 residues result in disruption of intra- and interhexamer interfaces, whereas mutations in the core of the hexamer are less likely to disrupt virion assembly (21).

Elite Controllers of HIV Preferentially Target Multiple Sites in Sector 3. Our results suggest that applying CTL pressure at multiple sites contained in sector 3 is likely to facilitate durable control of viral load to low levels; multiple mutations that allow immune escape are more likely to hurt viral fitness significantly, because multidimensional escape pathways are restricted. Can HLA (human MHC class I) molecules present peptides containing multiple sites within the identified vulnerable region? Because elite controllers achieve viral control (16), we first analyzed the locations of sites contained in peptides targeted in individuals with HLA alleles associated with spontaneous HIV control (HLA-A*25, B*57, B*27, B*14, Cw*08) (22). Remarkably, the sector with the largest proportion of sites contained in the dominant peptides targeted by controllers is sector 3 ($P = 3{\cdot}10^{-7}$; Fig. 3*A* and *SI Appendix*, Table S3). CTL pressure imposed by individuals with these HLA alleles, however, is not the driver of this collectively coevolving group of sites for reasons that include:

i) There are clear structural explanations (Fig. 2) for the collective correlations that characterize sector 3 sites, and structural features are independent of immune pressure.
ii) Negative correlations cannot be induced by immune pressure, because mutations induced in persons with the same HLA would be more likely to arise than if they occurred independently.
iii) As noted, a quasi-sector describes weak correlations associated with HLA-associated mutations (*SI Appendix*, Fig. S5), and these sites are not a part of sector 3.

Thus, the sector we identify to be most vulnerable by analyses of viral sequences and supramolecular structures is the most targeted by controllers.

Examining targeted epitopes lends further insights. For example, eight amino acids in the dominant epitopes targeted by B57-restricted CTLs are in sector 3; this multiplicity is further augmented by the enhanced cross-reactivity of these CTLs (23). $B14^+$ individuals target three amino acids in this sector. Population studies indicate that HLA-$B57^+$/$B14^+$ persons are partic-

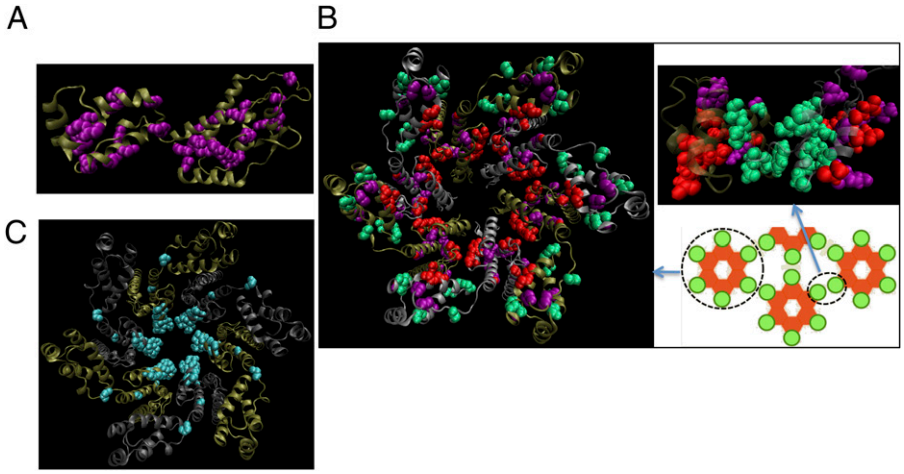

Fig. 2. Protein structures reveal the origin of collective correlations. (*A*) Sector 3 sites represented on the structure of the p24 monomer (PDB ID code 3GV2) are shown as purple spheres. (*B*) Sector 3 sites represented on the structure of the p24 hexamer (PDB code 3GV2) and the structure of the interface between hexamers (PDB code 2KOD). Sites at interfaces between two p24 proteins belonging to two adjacent hexamers are shown in green, and sites at interfaces between two p24 molecules within a hexamer are shown in red. The few remaining sites in sector 3 that are not part of these interfaces are shown in purple. (*C*) Sector 1 sites are shown in cyan on the structure of the p24 hexamer.

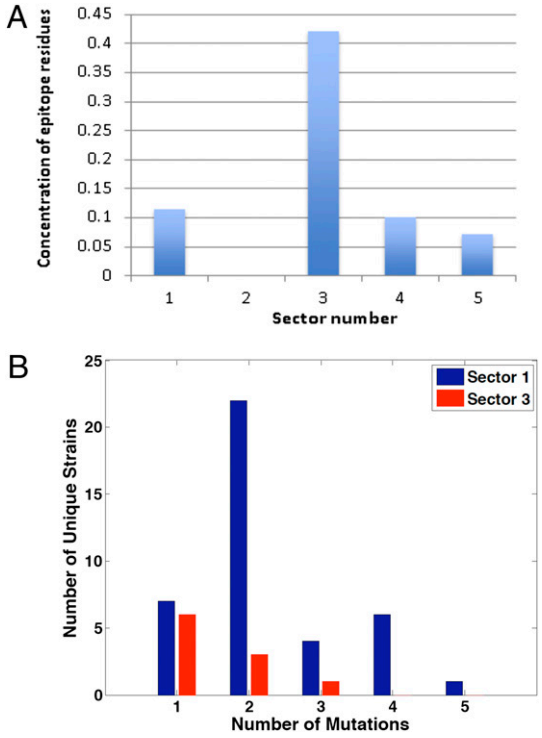

Fig. 3. Concentration of sites from the dominant epitopes presented by elite controllers in each sector and comparison of mutation patterns in sectors 1 and 3 in viruses derived from elite controllers. (*A*) Epitope is defined as dominant if it is the most targeted (in Gag) by individuals with the corresponding HLA allele. Concentration is the number of epitope sites in the sector divided by the total number of residues in the sector. The *P* value of association with each sector is in *SI Appendix* 10. Only sector 3 is significantly enriched with sites from the dominant epitopes of HLA molecules associated with control. Sectors 4 and 5 are not targeted at multiple points because they contain only 10 and 14 sites, respectively. Sectors 1 and 3 are targeted at multiple points. (*B*) Comparison of the number of unique viral strains obtained from a cohort of elite controllers that contains different numbers of mutations in sectors 1 and 3.

ularly efficient controllers of HIV infection. Our results suggest that this is because viral strains that can escape the multiple points of immune pressure in this sector are likely to have very low fitness. Thus, point mutants with lower viral fitness that only partially escape immune pressure or strains with multiple mutations that are not negatively correlated are likely to be the only options for the virus. This is consistent with the observed lower viral load and lower average viral fitness in controllers (16). Note also that CTLs restricted by HLA molecules associated with progression do not target sector 3 sites significantly ($P = 0.37$; *SI Appendix*, Table S3).

Virus Sequences from Patients Show That Multiple Mutations in Sector 3 Are Unlikely. Roughly 10% of sites in sectors 1, 4, and 5 are targeted by controllers, and sector 2 is not targeted (Fig. 3*A*). However, because sectors 4 and 5 are composed of only 10 and 14 sites, respectively, only 1 site (and epitope) is targeted by controllers in these sectors. Because sector 1 comprises 79 sites, many sites in sector 1 are targeted by controllers. If two regions of the proteome are equally targeted, all things being equal, the number and types of mutations observed in these regions should be similar. Sectors 1 and 3, both in Gag, exhibit similar levels of single site conservation, but we find that sector 3 is more multidimensionally constrained. Thus, we predict that even though sector 3 is targeted more than sector 1, fewer HIV strains with multiple mutations in sector 3 sites would be viable compared with those with similar mutations in sector 1.

To test this prediction, we sequenced viruses obtained from elite controllers because they target both sectors 1 and 3. We obtained 72 sequences from plasma and 103 including both plasma and peripheral blood mononuclear cells (PBMCs). The qualitative results obtained from analyzing either set of sequences are the same (Fig. 3*B* and *SI Appendix* 18). As per our predictions, the frequency of single mutations in sectors 1 and 3 is similar but multiple mutations are significantly rarer in sites that comprise sector 3. Our results suggest that this is because viral strains that can escape multiple points of immune pressure in sector 3 are more likely to be associated with negative correlations, and hence have particularly low replicative fitness because of defective capsid assembly. Thus, they are unlikely to be viable, and we do not observe them. These results show that the calculated differences between collective correlations among sites that comprise sector 3 and other sectors (Fig. 1) result in qualitative differences in viral mutations observed in humans that target sector 3 sites. Note also that almost all the few viral strains that we observe with multiple mutations in sector 3 are associated with positive correlations (*SI Appendix*, Fig. S13). Although only positive correlations can correspond to compensatory mutations, positive correlations can also correspond to multiple mutants that are only moderately less fit compared with single mutants, unlike negative correlations, which are only associated with multiple mutants that are very deleterious.

Our results strongly support our hypothesis that applying CTL pressure at multiple points in a multidimensionally constrained group of sites can trap the virus, because we find that controllers target multiple sites in sector 3, but strains with multiple mutations in sector 3 are not viable. Thus, viable strains are still likely to be subject to immune pressure.

Immunogens That May Induce CTL Responses in a Population That Hurt HIV. Can such potent responses be induced by vaccination in persons who are not blessed with HLA alleles that mount the correct dominant responses? Individuals with other HLA alleles present subdominant peptides containing sites in the vulnerable regions we have identified (using the Epitope Tables of the Los Alamos HIV Molecular Immunology Database, http://www.hiv.lanl.gov/content/immunology/tables/tables.html). This suggests the following strategy for design of immunogens that could induce memory CTL responses targeting the most vulnerable regions of HIV in a population: Select protein segments that simultaneously maximize sites in sector 3 (and some sites in sector 1), which are characterized by negative correlations, and those contained in subdominant and dominant epitopes presented by HLA molecules that span a broad section of a population. Such immunogens should elicit memory CTLs that only target the multidimensionally conserved regions of HIV, because epitopes that are dominantly targeted ineffectually are excluded. During natural infection, these memory CTLs are expected to mount robust responses that hit HIV where it hurts early, before naive CTLs mount ineffective dominant responses. Mounting the right responses early is important for HIV infections (2), because there is a short window of time before the immune system is compromised.

To illustrate this strategy, we considered a target population, white Americans with the 25 most frequent haplotypes (~39% of this population) (24). We create all groups of 10 known epitopes (as an example) presented by HLA-A/B molecules that comprise these haplotypes and select the top 10 groups according to the two criteria noted above (*SI Appendix* 12). One top-scoring immunogen (p24 residues 160–188 and 240–277) contains at least 1 or 2 targeted epitopes in 97% and 56% of the target population, respectively (a lower bound, because many epitopes are unknown). Previous empirical data show that controllers target p24 sites 240–272 (25). Sites 160–188 contain only 1 epitope tar-

geted by known controllers but represent an equally vulnerable region that provides broad coverage.

Discussion

Immunogens designed in this way to include only the multidimensionally constrained regions of the HIV proteome, which also contain epitopes presented by diverse HLAs in a population, are candidates for peptide vaccines using synthetic vectors (26), or by linking each segment together, they can be analogs of mosaic immunogens (27) delivered by traditional vectors. Such an immunogen would target the most vulnerable regions of HIV rather than the whole proteome. Targeting the latter is more likely to elicit responses from which HIV can escape via mutations (28), while hindering a focused response directed at regions of immunological vulnerability we have identified. Our goal of identifying such regions, and the methods we have used toward this end, are different from previous efforts to study coevolution of HIV proteins (e.g., 14, 29–31). Further analysis following the logic described by us should reveal additional regions of HIV proteins that are multidimensionally constrained, thereby enhancing the list of regions that are candidates for inclusion in a potent vaccine. A practical vaccine may also require inclusion of CD4 epitopes and flanking residues needed for antigen processing which do not contain protein segments that are likely to elicit ineffectual memory CTL responses. In vitro experiments and studies with animal models are required to develop this new concept, which might also be applied to design efficacious immunogens against other viruses. Our results also suggest the design of new small-molecule inhibitors of HIV replication.

Methods

HIV Sequences and Similarity Analysis. Multiple sequence alignments of nucleotide sequences of HIV-1 Gag, RT, and Nef were downloaded from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/) (clade B, one sequence per patient). We checked the phylogenetic homogeneity of each set by performing a principal component analysis of the similarity matrix of sequences (*SI Appendix* 5).

Identification of HIV Sectors. The correlation matrix for a given set of sequences is cleaned from phylogeny and noise using RMT (as described in the main text and *SI Appendix* 1–4). The definition of sectors follows the strategy proposed by Halabi et al. (12), grouping positions with a particular spectral signature into a sector (*SI Appendix* 6).

Targeting by Elite Controllers. We defined a set of epitopes dominantly targeted by HLA alleles associated with control (22), using a published dataset of the frequency of recognition of epitopes in an HLA-specific population (32).

Sequencing Viruses from Elite Controllers. HIV-1 Gag was amplified using nested RT-PCR from previously frozen PBMCs or plasma from elite controllers as previously described (33). PCR products were purified, and cycle-sequencing reactions were performed using 60 HIV-1–specific sequencing primers. Population sequences were obtained using an ABI 3730 PRISM (Applied Biosystems) automated sequencer. Gag mutations in the sequences derived from elite controllers were defined with reference to the clade B consensus sequence.

ACKNOWLEDGMENTS. We thank D. Barouch and H. Eisen for valuable discussions. Financial support was provided by the Ragon Institute, a National Institutes of Health Director's Pioneer Award (to A.K.C.), National Institutes of Health Grants RO130914 (to B.D.W.) and PO1 AI074415 (to M. Altfeld and T.M.A.), The Howard Hughes Medical Institute (B.D.W.) and the Mark and Lisa Schwartz Foundation (B.D.W.) This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract HHSN261200800001E.

1. Walker BD, Burton DR (2008) Toward an AIDS vaccine. *Science* 320:760–764.
2. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF (2010) The immune response during acute HIV-1 infection: Clues for vaccine development. *Nat Rev Immunol* 10:11–23.
3. Amanna IJ, Slifka MK (2011) Contributions of humoral and cellular immunity to vaccine-induced protection in humans. *Virology* 411:206–215.
4. Barouch DH, Korber B (2010) HIV-1 vaccine development after STEP. *Annu Rev Med* 61:153–167.
5. Hansen SG, et al. (2009) Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nat Med* 15:293–299.
6. Hartl DL, Clark AG (2007) *Principles of Population Genetics* (Sinauer, Sunderland, MA), 4th Ed.
7. Altfeld M, Allen TM (2006) Hitting HIV where it hurts: An alternative approach to HIV vaccine design. *Trends Immunol* 27:504–510.
8. Goulder PJ, Watkins DI (2004) HIV and SIV CTL escape: Implications for vaccine design. *Nat Rev Immunol* 4:630–640.
9. Wigner EP (1967) Random matrices in physics. *SIAM Rev* 9:1–23.
10. Plerou V, et al. (2002) Random matrix approach to cross correlations in financial data. *Phys Rev E Stat Nonlin Soft Matter Phys* 65:066126.
11. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Phys Rev Lett* 83:1467–1470.
12. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138:774–786.
13. Bhattacharya T, et al. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* 315:1583–1586.
14. Brumme ZL, et al. (2009) HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS ONE* 4:e6687.
15. Kiepiela P, et al. (2007) CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med* 13:46–53.
16. Miura T, et al. (2009) HLA-associated viral mutations are common in human immunodeficiency virus type 1 elite controllers. *J Virol* 83:3407–3412.
17. Hoffman NG, Schiffer CA, Swanstrom R (2003) Covariation of amino acid positions in HIV-1 protease. *Virology* 314:536–548.
18. Ganser-Pornillos BK, Yeager M, Sundquist WI (2008) The structural biology of HIV assembly. *Curr Opin Struct Biol* 18:203–217.
19. Pornillos O, et al. (2009) X-ray structures of the hexameric building block of the HIV capsid. *Cell* 137:1282–1292.
20. Troyer RM, et al. (2009) Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog* 5:e1000365.
21. Adamson CS, Jones IM (2004) The molecular basis of HIV capsid assembly—Five years of progress. *Rev Med Virol* 14:107–121.
22. Pereyra F, et al.; International HIV Controllers Study (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330:1551–1557.
23. Kosmrlj A, et al. (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* 465:350–354.
24. Maiers M, Gragert L, Klitz W (2007) High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* 68:779–788.
25. Streeck H, et al. (2007) Recognition of a defined region within p24 gag by CD8+ T cells during primary human immunodeficiency virus type 1 infection in individuals expressing protective HLA class I alleles. *J Virol* 81:7725–7731.
26. Melief CJ, van der Burg SH (2008) Immunotherapy of established (pre)malignant disease by synthetic long peptide vaccines. *Nat Rev Cancer* 8:351–360.
27. Barouch DH, et al. (2010) Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nat Med* 16:319–323.
28. Altfeld M, Goulder PJ (2011) The STEP study provides a hint that vaccine induction of the right CD8+ T cell responses can facilitate immune control of HIV. *J Infect Dis* 203:753–755.
29. Liu Y, Eyal E, Bahar I (2008) Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* 24:1243–1250.
30. Fares MA, Travers SAA (2006) A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* 173:9–23.
31. Bickel PJ, et al. (1996) Covariability of V3 loop amino acids. *AIDS Research and Human Retroviruses* 12:1401–1411.
32. Streeck H, et al. (2009) Human immunodeficiency virus type 1-specific CD8+ T-cell responses during primary infection are major determinants of the viral set point and loss of CD4+ T cells. *J Virol* 83:7641–7648.
33. Miura T, et al. (2008) Genetic characterization of human immunodeficiency virus type 1 in elite controllers: Lack of gross genetic defects or common amino acid changes. *J Virol* 82:8422–8430.