Supporting Information

Dahirel, V. et. al. , 'Coordinate linkage of HIV evolution reveals regions of immunologic vulnerability' (2011)

1. Definition of the correlation matrix

For each position *i* of a given protein or polyprotein, the type of amino acid in each sequence *s* of the Multiple Sequence Alignment (MSA) is represented by a binary variable $x_i(s)$, where $x_i(s) = 1$ if the amino-acid is the most frequent amino acid at this position within the MSA, and $x_i(s) = 0$ if it is another amino acid (binary representation). Since different positions have different levels of conservation (standard deviation of $x_i(s)$), we define a normalized cross-correlation matrix **C**, with elements:

$$C_{ij} = \frac{\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s}{\sqrt{V_i V_j}} \quad \text{(S1)}$$

Here, $V_i \equiv \langle x_i^2 \rangle_s - \langle x_i \rangle_s^2$ is the variance of the binary variable x_i. Such a matrix **C** represents the extent of pairwise coupling between mutations at different positions in a protein segment. In the case of Gag, a heat map representation of this matrix **C** is shown in Fig. S1.

An alternate means to represent the extent of coupling between mutations at different positions in the polyprotein is to follow the formalism of Statistical Coupling Analysis (SCA) introduced by Halabi et. al.[1], which uses the Statistical Coupling Matrix defined as,

$$C_{ij}^{SCA} = \phi_i \phi_j \left| \langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s \right| \quad \text{(S2)}$$

where $\phi_i \equiv \ln\left[\frac{\langle x_i^s \rangle_s (1 - q^{a_i})}{q^{a_i}(1 - \langle x_i^s \rangle_s)}\right]$, and q^{a_i} is the frequency of the most frequent amino acid a_i at position *i* averaged over all positions and over all sequences. Analysis using this correlation matrix is discussed in S9.

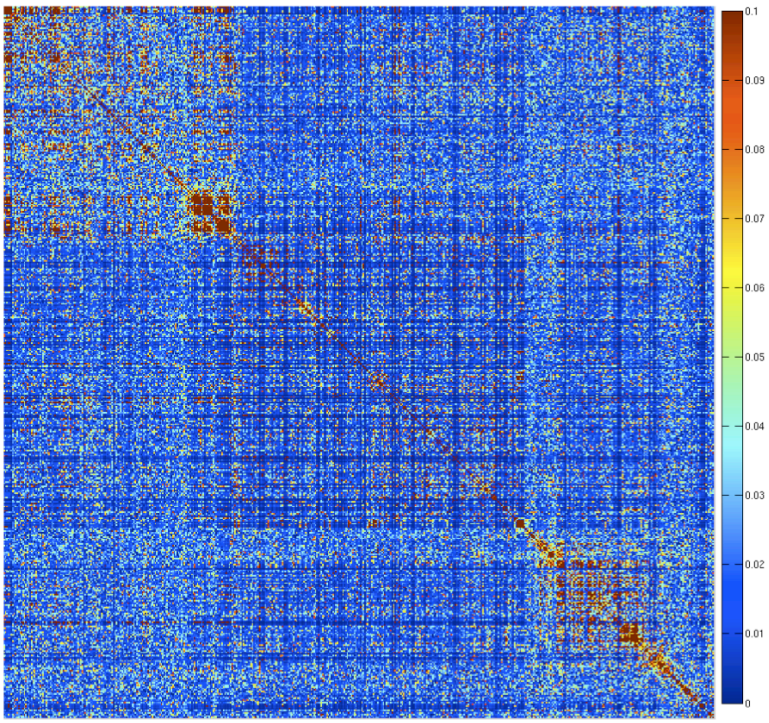

Figure S1. Uncleaned correlation matrix for Gag.

2. Use of Random Matrix Theory (RMT) to clean the correlation matrix

Consider two completely independent random variables x_i and x_j. If one computes the correlation coefficient C_{ij} between these two variables using a finite sample of values of x_i and x_j, then the result will be a finite non-zero number. More generally, the Correlation Matrix **R** between N independent random variables computed with M values of those variables has finite non-zero elements, R_{ij}. The values of R_{ij} do not reflect any real correlations between x_i and x_j, since these variables are strictly independent. However, random matrices such as **R** have well-defined statistical properties, which can be used as a reference to detect real correlations. In particular, it can be shown that, for correlation matrices defined by eq. S1 in the limit $N \to \infty$ and $M \to \infty$, with $Q \equiv M/N$ fixed, the probability density function of eigenvalues λ_i is non-zero only for values of λ_i between the bounds $\lambda_- \le \lambda_i \le \lambda_+$. λ_- and λ_+ are the lower and upper bounds on the eigenvalues of the random matrix **R**, given by[2]

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \quad \text{(S3)}$$

For finite M and N, the abrupt cut off of the probability density function turns into a fast decaying edge.

As RMT describes spectral properties, the correlation matrix **C** can be written as its eigenvalue decomposition in the Dirac notation[3]:

$$C = \sum_{k=1}^{N} \lambda_k |k\rangle\langle k| \quad \text{, or}$$

$$C = \sum_{\lambda_k < \lambda_-} \lambda_k |k\rangle\langle k| + \sum_{\lambda_- \le \lambda_k \le \lambda_+} \lambda_k |k\rangle\langle k| + \sum_{\lambda_k > \lambda_+} \lambda_k |k\rangle\langle k| \quad \text{(S4)}$$

Here, $|k\rangle$ is the kth eigenvector. The first sum in Eq. S4 is usually negligible, and the second term is dominated by noise-induced correlations. Therefore, the noise-cleaned correlation matrix $C_{cleaned}$ reads

$$C_{cleaned} = \sum_{\lambda_k > \lambda_+} \lambda_k |k\rangle\langle k| \quad \text{(S5)}$$

RMT bounds are defined for infinite M and N. In order to assess finite size effects, the distribution of eigenvalues for independent random variables can be determined numerically. For instance, in the case of a MSA represented by a set of binary variables, one can remove the real correlations between x_i and x_j by randomly permuting the values of $x_i(s)$ across different sequences *s*. For example, if there were only 3 sequences in the MSA, and a particular position had A, F, L as the amino acids in the 3 sequences, one random permutation would be F, L, and A. For our study of the 500 amino acid polyprotein Gag, 1600 sequences were used, giving theoretical values[2] (infinite size limit, Eq. S3): $\lambda_- \approx 0.198$ and $\lambda_+ \approx 2.42$. We performed 1000 permutations at each position to obtain 1000 randomized alignments. The distribution of the 500 000 obtained eigenvalues is shown in Fig. 1A (bottom spectrum). The higher edge of the distribution is (as expected) somewhat different than the asymptotic value of λ_+ = 2.42, but there is no eigenvalue above 3. Therefore, information contained in eigenvalues less than 3 are considered to be corrupted by noise.

The eigenvalue spectrum of the matrix **C** obtained for the real alignment of Gag is shown in Fig. 1A (top spectrum). Fourteen eigenvalues are higher than the noise threshold of 3. Those eigenvalues contain information corresponding to non-random correlations between the positions.

The principle of Random Matrix Theory can be similarly applied to the SCA correlation matrix[1], although in that case, no analytical formula for the bounds of the eigenspectrum is available. Since the properties of Random SCA Matrices are not as well defined as those of Random Correlation Matrices, we decided to analyze the data with the latter method. We checked that the results were qualitatively similar using the SCA correlation matrix (see S9).

3. The spectral contribution from phylogeny

The phylogenetic relationship[4] between the different sequences of the MSA can also give rise to correlations.

To understand this, let us consider the evolution of a hypothetical protein with *N* amino acids that mutate independently of each other. We further assume that the binary approximation holds, wherein each site on the protein may take on one of two values (1 and 0). Let $P_1^i(t)$ and $P_0^i(t)$ represent the probabilities that the residue occupying site *i* is '1' and '0' respectively at time *t*. Using the conservation relation $P_1^i(t) + P_0^i(t) = 1$, we can describe the time evolution of the probability $P_1^i(t)$ for each site by the master equation,

$$\frac{dP_1^i(t)}{dt} = -\mu_{1->0} P_1^i(t) + \mu_{0->1}(1 - P_1^i(t)) \qquad i = 1,2,3,....,N \qquad \text{(S6)}$$

Here, $\mu_{1->0}$ and $\mu_{0->1}$ are the mutation rates from state '1' to state '0' and vice-versa. These rates are the same for all *N* positions because of our assumption that the mutations at each position are independent. We assume an initial condition $P_1^i(0)$ = 1 for all sites *i*. In this simple toy-model, the uniform initial condition represents phylogeny. The solution to Eq. S6 is,

$$P_1^i(t) = \frac{\mu_{0->1}}{\mu_{0->1} + \mu_{1->0}} + \frac{\mu_{1->0}}{\mu_{0->1} + \mu_{1->0}} e^{-(\mu_{0->1} + \mu_{1->0})t} \qquad \text{(S7)}$$

Here, for simplicity, we consider that we have an infinite number of sequences sampled uniformly between the times t=0, and t=T. We compute the correlation function between two positions *i* and *j*

$$C_{ij}(T) = \frac{1}{T}\int_0^T P_1^i(t) P_1^j(t) dt - \left[\frac{1}{T}\int_0^T P_1^i(t) dt\right]\left[\frac{1}{T}\int_0^T P_1^j(t) dt\right] \qquad \text{(S8)}$$

Since we consider all sites to be independent, $P_1^i(t)$ and $P_1^j(t)$ in Eq. S8 have identical expressions. Substitution of Eq. S7 in Eq. S8 and a little algebra yields,

$$C_{ij}(T) = \frac{\mu_{1->0}^2}{(\mu_{0->1} + \mu_{1->0})^3}\left[\frac{1}{2T} - \frac{1}{(\mu_{0->1} + \mu_{1->0})T^2}\right] + \frac{2\mu_{1->0}^2}{T^2(\mu_{0->1} + \mu_{1->0})^4} e^{-(\mu_{0->1} + \mu_{1->0})T} \qquad \text{(S9)}$$

$$- \frac{\mu_{1->0}^2}{(\mu_{0->1} + \mu_{1->0})^3}\left[\frac{1}{2T} + \frac{1}{(\mu_{0->1} + \mu_{1->0})T^2}\right] e^{-2(\mu_{0->1} + \mu_{1->0})T}$$

As $T \to \infty$, $C_{ij}(T) \to 0$, but $C_{ij}(T)$ is non-zero for finite times T. Therefore, despite mutations at each site occurring independently, the presence of a common initial condition (i.e. phylogeny) contributes to a non-zero correlation between positions if we consider sequences that have evolved for finite times. If we sample sequences that have evolved for infinitely long times, however, this correlation vanishes as one would expect as the sequences have lost memory of the ancestral sequence – i.e, they are no longer phylogenetically related.

That phylogeny will contribute to a coherent mode in the correlation matrix can be illustrated using this simple calculation. Consider the N x N correlation matrix **C** for the hypothetical protein of chain length *N*. If each site evolves independently of others, **C** is a matrix of identical elements, each of which is equal to $C_{ij}(T)$ in Eq. S9. **C** has rank unity and its only non-zero eigenvalue is $NC_{ij}(T)$. This is a consequence of the Perron-Frobenius theorem in Linear Algebra[5]. The corresponding eigenvector is $\frac{1}{\sqrt{N}}(1,1,1,\ldots,1)$, which is a coherent mode.

4. Cleaning of Phylogeny

In practice, for the Gag, Nef, and RT alignments, the eigenvector corresponding to the highest eigenvalue is a coherent mode that largely reflects correlations due to phylogeny. Following the results from the previous section, the simplest way to remove phylogenetically induced correlations is to remove the contribution coming from this coherent eigenvector in the sum (Eq. S5). However, since the highest eigenvalue influences the remaining eigenvalues, it would be preferable to have a way to recompute the eigenspectrum after removal of the highest eigenvalue. Such a method is outlined in a study by Plerou et. al.[2].

Phylogeny can be considered as a field $M(s)$ that is coherently applied to each position, but depends on the sequence *s*. We can then statistically write the binary variable $x_i(s)$ as

$$x_i(s) = \alpha_i + \beta_i M(s) + \varepsilon_i(s) \qquad \text{(S10)}$$

where $\langle \varepsilon_i(s) \rangle = 0$ and $\langle \varepsilon_i(s) M(s) \rangle = 0$. The phylogenetic field can be approximated by the contribution of the first (highest) eigenvalue to the value of the variables $x_i(s)$:

$$M(s) = \sum_{i=1}^{N} u_i^1 x_i(s) \qquad \text{(S11)}$$

where the coefficients u_i^1 are the components of the eigenvector corresponding to the highest eigenvalue.

The value of the parameters α_i and β_i are estimated through a least square regression, replacing $M(s)$ by its approximation (Eq. S11) in the expression (Eq. S10) of the variables, $x_i(s)$. We then obtain a MSA represented by the variable, $y_i(s)$,

$$y_i(s) = \alpha_i + \varepsilon_i(s) \qquad \text{(S12)}$$

Notice that the value of α_i does not influence the correlation matrix **C**.

In the case of the Gag alignment, the correlation matrix of the variables $y_i(s)$ is similar to that of the variable $x_i(s)$ after removal of the contribution due to the highest eigenvalue. This similarity is made apparent in the heat map representation of the cleaned correlation matrix (Fig. S2). In this heat map we kept the order of the rows (or columns) that was defined using the modified variables $y_i(s)$. We can see that the same sectors still appear vividly.

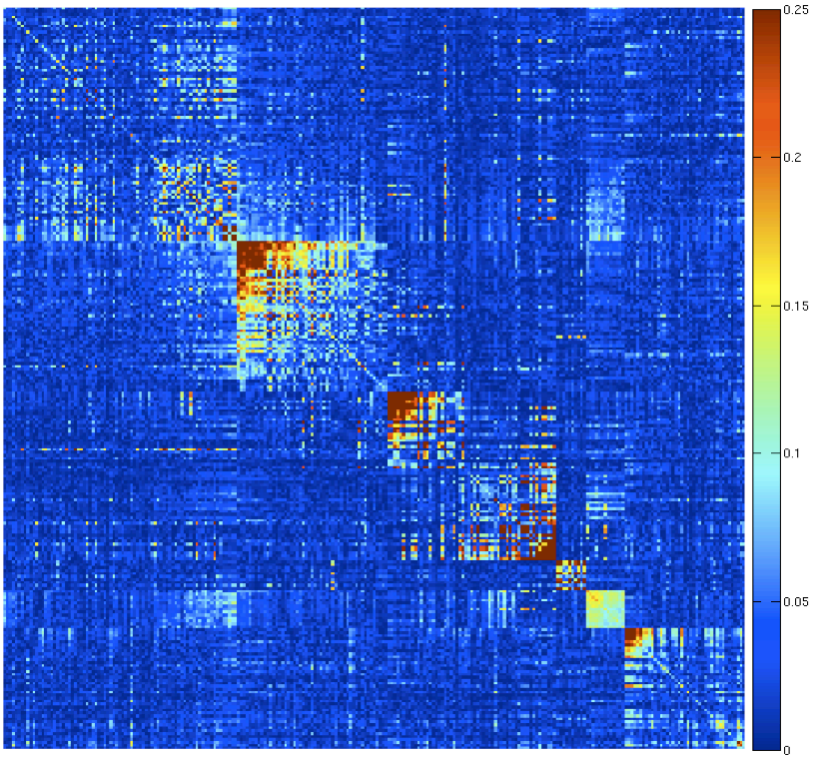

Figure S2. Cleaned correlation matrix for Gag obtained with an alternative procedure to clean phylogeny compared to that shown in Fig. 1B. In this case, we simply remove the contribution coming from the first eigenvalue. The ordering of positions is the same as the ordering defined to represent the cleaned correlation matrix is Fig. 1B. The sectors remain qualitatively similar (only the quantitative value of C_{ij} is different).

5. Removing evolutionarily distinct sequences

The method to clean the correlations induced by phylogeny only works when the sequences are not derived from different clades, or clusters that are well separated phylogenetically from each other.

The similarity between two sequences s and t can be measured as the coefficients of the similarity matrix $\Gamma_{st} = \langle x_i(s) x_i(t) \rangle_i - \langle x_i(s) \rangle_i \langle x_i(t) \rangle_i$. This measure can be used to detect groups of sequences that are evolutionarily separated from the rest of the sequences. This can be visualized by projecting sequences along the eigenvectors corresponding to the highest eigenvalues of the similarity matrix. In Fig. S3 , we show these projections for a set of 1788 clade B Gag sequences downloaded from the Los Alamos HIV database[6]. We clearly see 2 clusters for high contributions to the second eigenvector. Keeping the sequences in these clusters leads to phylogenetically induced correlations between positions at the end of the p6 protein. These correlations are not removed through the previously mention cleaning procedure. If we remove the small number (188) of sequences corresponding to the two small clusters with high contributions to the second eigenvector from the MSA, those spurious correlations disappear. All other aspects of our sector analyses remain qualitatively the same, however. In our study of Nef (S13) we also had to remove sequences to preserve the phylogenetic homogeneity of the MSA.

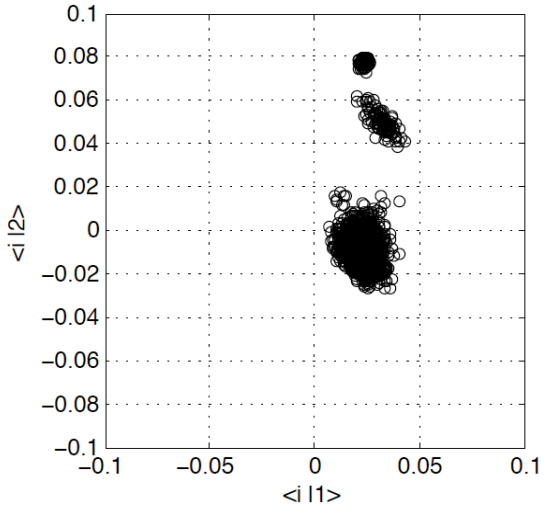

Figure S3. Projections of the components of the 2 eigenvectors corresponding to the 2 largest eigenvalues of the similarity matrix for the Gag MSA. The sequences within the two clusters corresponding to high projections along the second eigenvector are removed from the MSA before computing the correlation matrix.

6. Definition of sectors

Once the matrix is cleaned of noise and phylogeny, the remaining coefficients reveal information about real correlations between positions. If the correlation coefficient is strong (in absolute magnitude), it means that the mutations at each position are correlated, while if this coefficient is small, it means that evolution at both positions is roughly independent. Halabi et. al.[1] argue that some positions could be clustered together into sectors, as companies can be associated into economic sectors based on the correlation between the prices of their stocks[2] (see main text).

This clustering can be uncovered by an examination of the position-wise loadings on eigenvectors corresponding to the highest eigenvalues of **C**. Indeed, if two positions have high loadings along one of these eigenvectors $|k\rangle$, then the term $\langle i|k\rangle\langle k|j\rangle$ is a strong contributor to the coefficient C_{ij}. Then each eigenvector reflects a collective correlation in the evolution of different positions. Collectivity should not be important for immune or drug induced evolution, which involves a few residues, and only acts within a small subset of the strains. Therefore, the correlations that define a sector are more likely due to the functions of the protein. Moreover, in the cases we have studied, the proteins are not subject to any drug pressure (RT sequences are taken from drug naïve individuals), and the MSA contains sequences obtained from individuals with various HLA haplotypes. More discussion of immune pressure – induced correlations is provided in S8.

7. Determination of the sectors of Gag

The projections of the loadings (values of the components) for the eigenvectors corresponding to the highest eigenvalues were used to define Gag sectors (Fig. S4).

Three groups of sites have a strong contribution to the first eigenvectors $|1\rangle$ to $|3\rangle$, respectively, leading to the definition of three sectors (sectors 1-3). The small group of residues with large positive loadings along eigenvector $|2\rangle$ does not show strong correlations with the residues that possess large negative loadings along the same eigenvector (Fig S4A). This group of 14 residues constitutes a distinct sector (sector 5). Indeed, those two sectors (2 and 5) are defined by directions that are almost orthogonal in the map defined by eigenvectors $|1\rangle$ and $|2\rangle$ (Fig S4A). Also, one sector contains residues that have strong coefficients along eigenvector $|5\rangle$ (sector 4). Although these sectors are defined according to a particular eigenvector, they can also have strong contributions along other eigenvectors. In the case of sector 3, the sector residues have significant loadings along eigenvectors $|4\rangle$ and $|5\rangle$ with positive and negative signs (Fig S4B and S4C), which is the collective signature of the strong anti-correlations within this particular sector (see main text).

Defining the borders of the sectors is not clear-cut. When residues have high contributions to several eigenvectors, we chose to prevent sector overlap by assigning the residue to the sector with which this residue has its highest mean absolute correlation coefficient. Moreover, we chose to exclude from the sectors

those residues that did not show any correlation higher than 0.1 (in absolute magnitude) with any other sector residue. This limit is based on the distributions of the values of coefficients from random matrices. For the random matrices, we find that correlation coefficients with magnitude greater than 0.1 arise with probability $< 5.10^{-3}$.

The residues belonging to each sector are reported in table S1. Note that 289 residues do not fall in any sector.

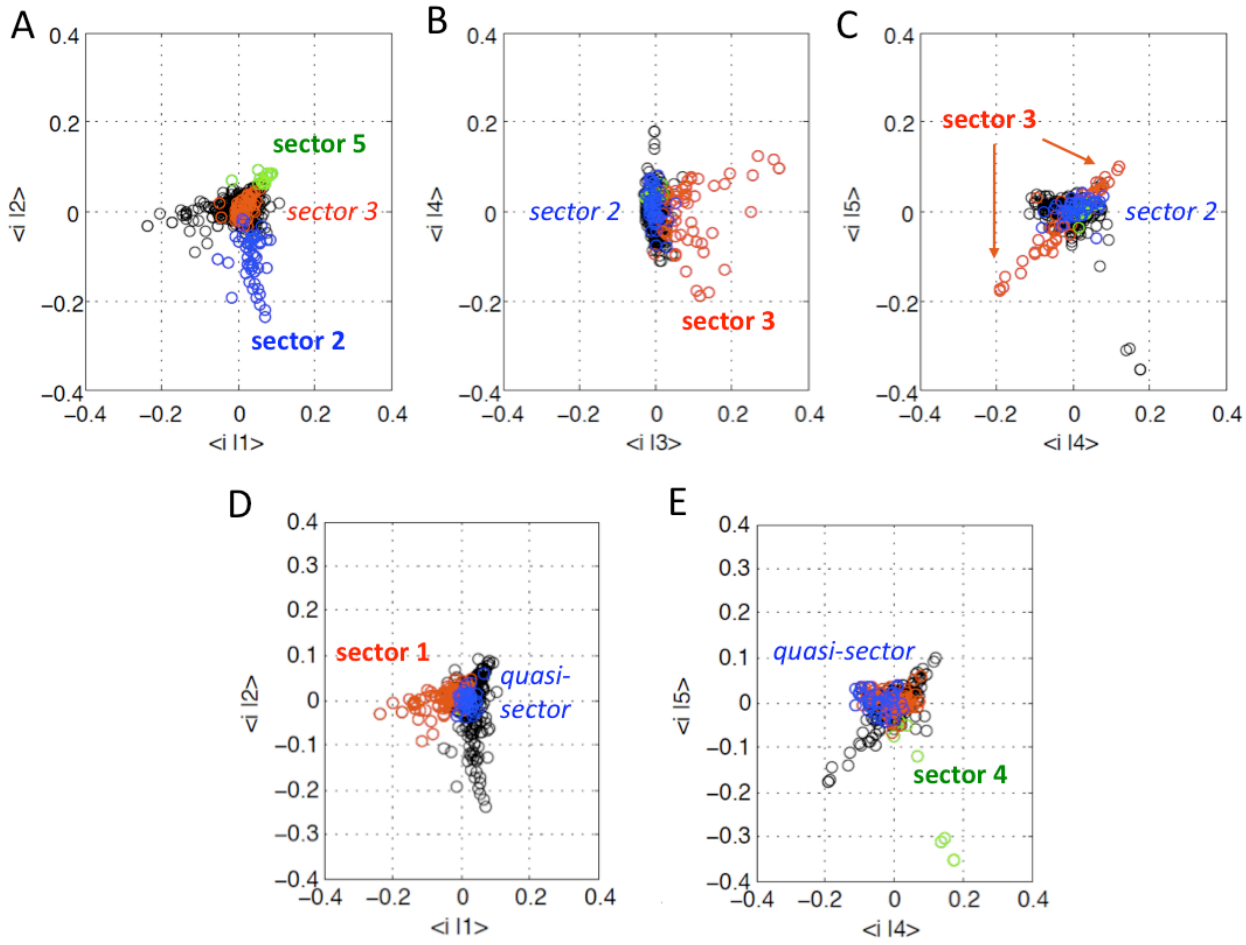

Figure S4. Projections of the components of the 5 eigenvectors corresponding to the 5 highest eigenvalues of the correlation matrix, after removal of phylogeny (see S4), for the Gag multiple sequence alignment. Each map shows the projection along 2 eigenvectors. The definition of sectors using these maps is described in S6 and above.

Table S1. Amino acids that comprise the sectors of Gag; numbers are those of the Clade B consensus sequence.

Sector 1		Sector 2		Sector 3		Sector 4	Sector 5	Q-Sector
1	88	23	446	53	305	166	17	18
2	94	37	450	140	306	197	31	30
3	97	178	452	163	310	211	47	54
4	99	379	455	167	316	222	137	62
5	100	381	457	169	317	236	161	69
6	108	386	459	170	319	237	261	90
8	118	391	461	171	323	308	275	125
9	120	392	462	172	326	318	278	130
11	122	393	463	174	338	354	290	146
12	123	394	464	175	344	396	298	147
14	128	395	489	179	345		324	159
16	129	399		180	346		334	173
19	131	400		181	347		337	176
20	133	402		182	363		343	200
21	134	405		185	364			218
24	135	406		186	365			219
27	136	407		187	366			223
29	138	408		189	367			224
32	139	412		191				228
33	141	413		198				230
35	142	414		199				234
36	143	416		212				242
38	144	417		221				248
39	145	419		225				252
41	148	420		229				255
45	149	423		233				256
48	150	430		240				264
50	151	431		243				267
51	152	432		245				268
52	153	434		249				273
57	154	435		257				280
60	155	437		260				281
63	156	438		263				286
73	158	439		265				312
77	160	440		269				341
79	251	442		284				357
83	276	443		288				362
86	279	444		291				374
87	433	445		295				375
								376
								401
								403

8. A quasi-sector induced by immune pressure

A careful investigation enabled us to distinguish a group of correlated residues that do not contribute to any eigenvector significantly; i.e., these residues are collectively only weakly coupled (see Fig. S4, panels D and E). This suggested to us that the correlations within this group, or quasi-sector, might be induced by immune pressure.

In order to determine which pairs of residues are likely correlated due to immune pressure, we need to know which mutations are seen in individuals carrying a given HLA allele. Then the presence of this allele would correlate those mutations. This information is available in a previous study by Brumme et. al.[7].

In the case of Gag, this study identifies 344 associations between a given HLA and a single site polymorphism. Assuming that all residues associated with the same HLA allele are potentially correlated because of immune pressure, we identified 358 possible pairwise associations. Those associations are superimposed on the sector map in Fig. S5. 155 identified pairs involve residues that are not part of any sector. None of those pairs involves a residue from sectors 3-5. Sector 1 contains one of such coupled pair, and sector 2 contains four of them. However, 41 associations are between positions within the quasi-sector. This supports our hypothesis that this group of sites forms a weakly coupled quasi-sector that reflects correlations that are induced by immune pressure. Interestingly though, other studies have shown that some individual pairs of residues within this quasi-sector are functionally correlated. For instance, it has been established that a mutation at position 173 restores fitness after a mutation at position 264, which is believed to be related to cyclophylin binding[8].

9. Representation of the cleaned correlation matrix

In Fig. 1B for Gag, Fig. S8 for Nef and Fig. S9 for RT the cleaned correlation matrices are represented by heat maps. However, the positions are grouped according to sites contained in each sector, not the linear sequence of the protein. First, the sector positions are grouped together. For instance in Gag, the first 79 positions are the residues belonging to sector 1, followed by the 51 positions that are of sector 2, and so on. Secondly, within each sector, the positions are ordered by decreasing value of their projection along the eigenvector with which they are defined. For instance, the first position of sector 1 has the highest value of $\langle i|1\rangle$. For the quasi-

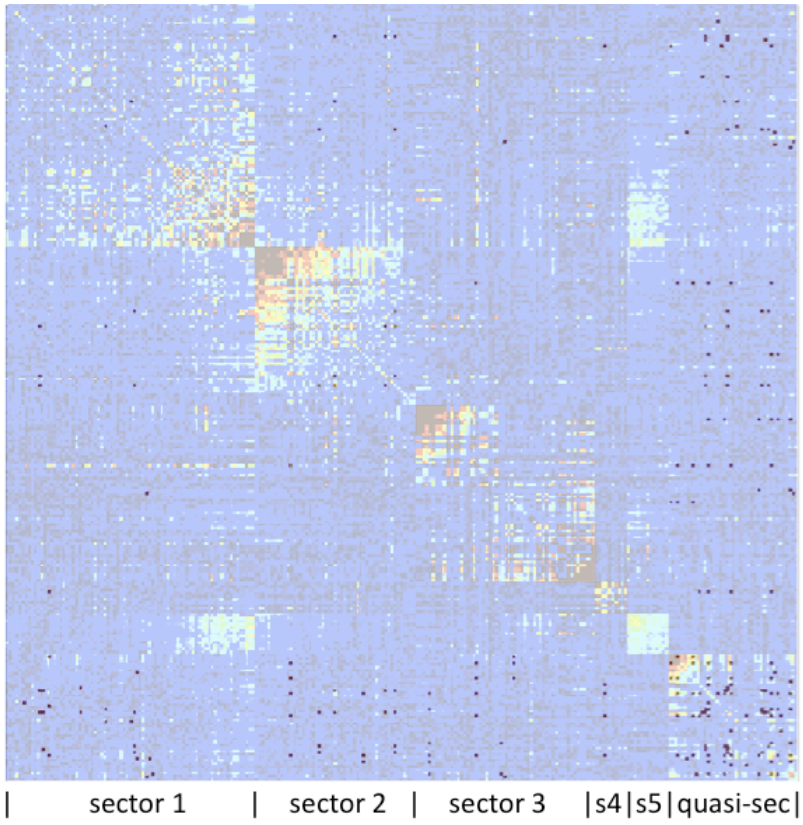

Figure S5. Map of pairs of positions potentially coupled by immune pressure (marked by dots), superimposed on the sector map.

sector, which is defined according to weak contributions to several eigenvectors, the ordering is derived from the loadings along the 7th eigenvector.

In figure S6, we also show the cleaned SCA correlation matrix represented using the same order. This figure reveals important intra-sector correlations and weak inter-sector correlations. Thus, the qualitative picture is consistent using the SCA matrix.

We have also determined the sectors using the SCA correlation matrix, and the results are qualitatively the same (not shown). The SCA sectors include more residues as some non-sector residues obtained using Eq. S1 become parts of sectors, but the sites in a given sector remain in the same sector regardless of whether we use Eq. S1 or Eq. S2. The small number (14) of residues that comprise sector 5 (using Eq. S1) became a part of sector 3 (using Eq. S2). However, this does not change any qualitative results since essentially the same proportion of sector 3 residues (~66 %) are associated with the interfaces important for capsid assembly (Fig. 2), regardless of whether we derive the sectors using Eq S1 or Eq S2.

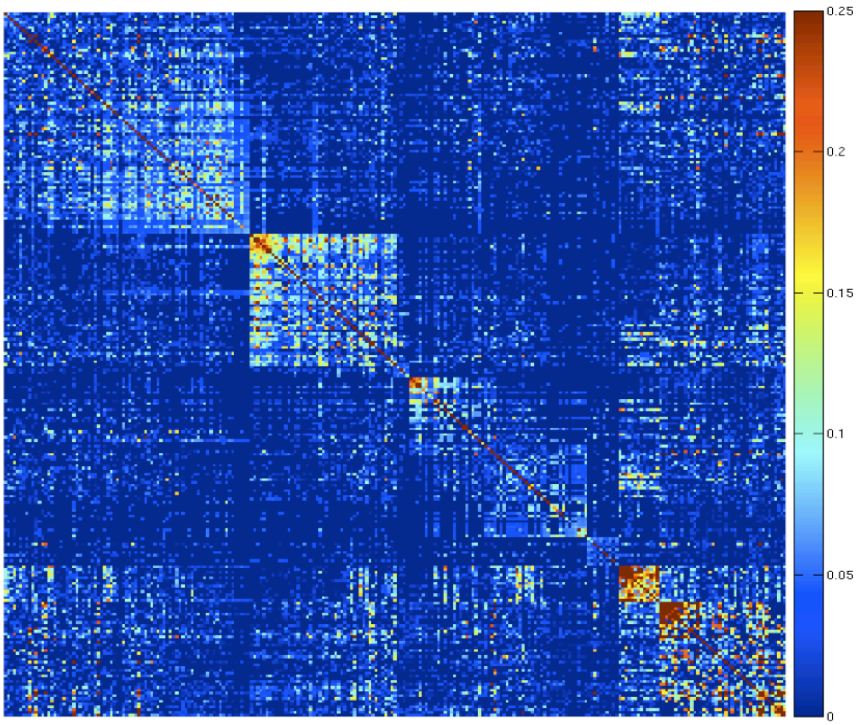

Figure S6. Cleaned SCA correlation matrix for Gag. The ordering of positions is the same as the ordering defined to represent the cleaned correlation matrix in Fig 2B. The sectors remain qualitatively similar.

10. Association between immune pressure in controllers and sectors

We wished to determine whether individuals with genes associated with superior control of HIV target sites that are contained in the sectors that we determined to be vulnerable to CTL pressure. To accomplish this, we used information from an extensive HLA allele/disease progression association study, published recently[9]. Moreover, we used immunodominance data from another study[10] as explained below.

The genome association data provides 'odds ratios' and 'p-values' of association between HLA alleles and viral control for a cohort of individuals from a European population, and also for a cohort of individuals from an African-American population. We considered only those alleles with an odds ratio higher than the odds ratio for B*14, or lower than the odds ratio for B*08, in the case of the European population to classify alleles as associated with control or progression, respectively. Among these alleles, we only kept those for which the frequency of recognition of some of their epitopes has been measured. All the chosen alleles have a p-value of association lower than 10^{-2}. Among those alleles, if they are present within the African American population, they are similarly associated with control, or progression. We obtained a group of 5 HLA alleles associated with control, HLA A*25, B*57, B*27, B*14, and Cw*08. Also, 4 alleles are associated with progression, HLA B*07, B*08, B*35, and A*29.

To determine which epitopes are significantly presented for each of these alleles, we referred to a set of measurements of the frequency of epitope recognition by T-cells from HIV-infected people carrying specific HLA alleles[10]. It is possible that this study misses some important epitopes, but it is the most extensive study that has been published so far. We picked epitopes according to 2 different criteria. A first set of epitopes contains the most frequently recognized epitope in acute phase for each HLA allele, and the most recognized epitope in chronic phase for each HLA allele. This group (group 1) includes 7 epitopes restricted by HLAs associated with control, and 5 epitopes restricted by HLAs associated with progression. A second set of epitopes contains all epitopes that are recognized by at least 20 percent of individuals with that HLA in acute phase. This group (group 2) includes 7 epitopes restricted by HLA associated with control, but only 3 epitopes restricted by HLA associated with progression (see Table S2).

In order to quantify how the residues from these epitopes are distributed among the different sectors of Gag, we carried out standard p-value calculations. Within the null hypothesis defining these p-values, the probability of a given residue to fall within a sector is proportional to the number of residues in this sector.

Table S2. List of the epitopes of each group. Group 1 contains the most frequently recognized Gag epitopes in acute and in chronic phase, while the group 2 contains all epitopes that are recognized by more than 20 percent of the population carrying the specified HLA allele.

		Epitope group 1	Epitope group 2
HLA alleles associated with control	A25	QW11	QW11, EW10
	B57	TW10, KF11	TW10, KF11, IW9
	B27	KK10	KK10
	B14	DA9	DA9
	Cw8	RV9, TL9	
HLA alleles associated with progression	A29	LY9	LY9
	B07	GL9	GL9
	B08	EI8	EI8
	B35	PY9, WF9	

We define two p-values. p_+ is the p-value of association of with the s^{th} sector :

$$p_+ = \sum_{k=Ne,s}^{Ne,t} {}^{Ns}C_k \, {}^{Nt-Ns}C_{Ne,t-k} \,/\, {}^{Nt}C_{Ne,t} \approx \sum_{k=Ne,s}^{Ne,t} {}^{Ne,t}C_k \left(\frac{N_s}{N_t}\right)^k \left(\frac{N_t - N_s}{N_t}\right)^{Ne,t-k} \qquad \text{(S13)}$$

p_- is the p-value of "non-association" with the s[th] sector :

$$p_- = \sum_{k=0}^{Ne,s} {}^{Ns}C_k \, {}^{Nt-Ns}C_{Ne,t-k} \,/\, {}^{Nt}C_{Ne,t} \approx \sum_{k=0}^{Ne,s} {}^{Ne,t}C_k \left(\frac{N_s}{N_t}\right)^k \left(\frac{N_t - N_s}{N_t}\right)^{Ne,t-k} \qquad \text{(S14)}$$

Here N_t is the number of residues in Gag ($N_t = 500$), N_s is the number of residue in the s[th] sector, $N_{e,t}$ is the total number of residues of the considered epitopes, $N_{e,s}$ is the number of those $N_{e,t}$ residues in the s[th] sector. The results are presented in the tables S3 and S4.

Table S3. p-value of association p_+, and "non-association" p_- of the residues from group 1 epitopes with the different Gag sectors. N_s is the number of residue in the s[th] sector, $N_{e,s}$ is the number of epitope residues in the s[th] sector.

		"Controllers"			"Progressors"		
Sector	N_s	$N_{e,s}$	p_+	p_-	$N_{e,s}$	p_+	p_-
1	79	9	0.78	0.33	7	0.56	0.61
2	51	0	1.0	$6.0\ 10^{-4}$	1	0.99	$5.3\ 10^{-2}$
3	57	24	$3.0\ 10^{-7}$	1.00	6	0.37	0.77
4	10	1	0.75	0.60	0	1.0	0.41
5	14	1	0.86	0.42	2	0.35	0.87
Quasi-Sector	42	8	0.22	0.88	6	0.16	0.93
non-sector	247	26	0.98	$3.3\ 10^{-2}$	22	0.53	0.59

11. Strength of negative correlations

Within the binary approximation, for a given value of the frequencies f_i and f_j of the most abundant amino-acid at two sites i and j, one can compute the minimum value of the correlation coefficient C_{ij}.

The minimum value is obtained when the frequency of the double mutant is equal to zero (possible if $f_i + f_j > 1$). In that case, one can easily show:

$$f_{ij} = f_i + f_j - 1 \qquad \text{(S15)}$$

and then,

$$C_{ij} = \left(f_i + f_i - 1 - f_i f_j\right)/\sqrt{V_i V_j} = -\left(1-f_i\right)\left(1-f_j\right)/\sqrt{f_i(1-f_i)/f_j(1-f_j)} = -\sqrt{(1-f_i)(1-f_j)/f_j f_i} \qquad \text{(S16)}$$

The absolute value of this number is smaller when the frequencies are larger, i.e. when the two residues are more conserved. For example, for $f_i = f_j = 0.95$, we get a minimum value of -0.052. As a comparison, in the case $f_i = f_j$, the maximum value of C_{ij} is 1.

Table S4. p-value of association p_+, and "non-association" p_- of the residues from group 2 epitopes with the different Gag sectors.

		"Controllers"			"Progressors"		
Sector	N_s	$N_{e,s}$	p_+	p_-	$N_{e,s}$	p_+	p_-
1	79	17	$4.3\ 10^{-2}$	0.98	3	0.80	0.39
2	51	0	1.0	$5.4\ 10^{-4}$	0	1.0	$6.1\ 10^{-2}$
3	57	16	$4.8\ 10^{-3}$	1.0	4	0.34	0.83
4	10	2	0.41	0.84	0	1.0	0.59
5	14	1	0.86	0.41	1	0.52	0.84
Quasi-sector	42	8	0.23	0.87	4	0.17	0.94
non-sector	247	26	0.99	$2.6\ 10^{-2}$	14	0.40	0.74

As a consequence, the strength of a negative correlation cannot be as high as the strength of positive correlations.

In all the cases we have considered as highly negatively correlated (C_{ij} < -0.03 in Fig. 1d), the value of C_{ij} is actually equal to this limit. It means that for those pairs, the correlation coefficient could not be more negative.

However, from a statistical point of view, the probability that the correlation coefficient between two residues reaches this limit for many such pairs, as is the case in sector 3, is quite high. As an illustration, let's consider 2 conserved residues, i.e $f_i = f_j = 0.96$, and a sample of 1000 sequences. It means that there are 40 mutants at site i and 40 mutants at site j over the 1000 sequences. If the two residues evolve

independently, the probability that there is no double mutant is given by a hypergeometric distribution. This probability is 0.44. So just looking at the value of C_{ij} cannot tell in that case that the two residues are really anticorrelated. However, once we look into higher order correlations, things become different. If now we consider 3 conserved residues, i.e $f_i = f_j = f_k = 0.96$, and a sample of 1000 sequences, the probability that there is no strain with two mutations at any of those three sites becomes 0.081.

This example shows that if one pair has a negative value of C_{ij}, one cannot conclude that the pair is anticorrelated, while negative correlations within a group can be proved to be significant.

Random Matrix Theory is thus particularly adapted to reveal such groups of anticorrelated residues. In the case of Gag, the collective correlations create an anticorrelated mode (Fig. S4 C), corresponding to eigenvalues that cannot have emerged from noise.

12. Immunogen design

Our study has shown that two Gag sectors are immunologically vulnerable to multiple points of CTL pressure (sectors 1 and 3).

We propose a way to design optimal immunogens that would elicit multiple points of immune pressure on these vulnerable regions. To illustrate our strategy, consider a target population that must be vaccinated using the immunogen. We chose to design a vaccine for the Caucasian American population as a proof of principle. The frequency of HLA haplotypes appearing in this population is available[11]. For each HLA allele within those haplotypes, potential epitopes are listed in the list of "best-defined" epitopes from the HIV database[12]. We extracted a list of 31 p24 epitopes for the 25 most frequent haplotypes. We did not look at very rare haplotypes for which the knowledge of the presented epitopes is very poor. We then constructed all possible groups of 5 (this number is an example – we have also used groups of 10 epitopes, see below) such epitopes. The protein segments formed by these groups of epitopes are candidate immunogens. We screened the groups according to the following four criteria:

i) The epitopes have the maximum number of positions in sectors 1 or 3.
ii) The number of significant negative correlations ($C_{ij} < -0.03$) within the set of epitopes is maximum.
iii) The number of significant positive correlations ($C_{ij} > 0.1$) within the set of epitopes is minimum.
iv) The immunogen has a good coverage of the population. This can be quantified by the fraction of the population that elicits a response to at least one (or two) epitope.

The first criterion is the most important, as it is a direct consequence of our findings. We therefore use this criterion for a first screen, to keep the ~ 0.2 % (311) of all

possible 5 epitope combinations from the set of 31 epitopes with the highest number of residues in sector 1 or in sector 3. For criteria (ii) and (iii), we chose the thresholds to keep less than 10 percent of the 311 combinations. The 3 criteria (i), (ii) and (iii) lead finally to the selection of 11 combinations of epitopes. We finally sorted those combinations according to the coverage of the population. The results are shown in Table S5.

We also modified the thresholds to make a larger list, of 178 epitopes combinations, in order to see whether we missed epitopes with a slightly lower score, but a much better population coverage.

We also repeated the calculation for combinations of ten epitopes, using the same procedure. The results are presented in Table S6.

Table S5. 11 combinations of 5 epitopes (Ep.). The numbers correspond to the HXB2 numbering of amino acid positions. The coverage (Cov.) is the frequency of haplotypes that are known to present at least one of those 5 epitopes, within the Hardy-Weinberg approximation, and among the 25 most frequent haplotypes in the Caucasian American population. The double coverage (DCov.) is the frequency of those haplotypes known to present at least two of those 5 epitopes.

Combination	Ep. 1	Ep. 2	Ep. 3	Ep. 4	Ep. 5	Cov.	DCov.
1	167-175	180-188	259-268	260-267	260-269	0.578	0.304
2	167-175	180-188	240-249	260-269	291-300	0.696	0.162
3	167-175	180-188	240-249	260-269	308-316	0.414	0.116
4	160-168	167-175	180-188	260-269	261-270	0.365	0.059
5	167-175	180-188	240-249	260-269	261-270	0.414	0.059
6	167-175	180-188	254-262	260-269	261-270	0.425	0.059
7	167-175	180-188	240-249	259-268	260-269	0.638	0.038
8	167-175	180-188	240-249	260-267	260-269	0.638	0.038
9	167-175	180-188	240-249	260-267	261-270	0.638	0.038
10	167-175	180-188	240-249	260-269	306-316	0.529	0.038
11	167-175	180-188	240-249	260-267	306-316	0.723	0

Table S6. 11 combinations of 10 epitopes (Ep.). The numbers correspond to the Gag HXB2 numbering of amino acid positions. The coverage (Cov.) is the frequency of haplotypes that are known to present at least one of those 10 epitopes, within the Hardy-Weinberg approximation, and among the 25 most frequent haplotypes in the Caucasian American population. The double coverage (DCov.) is the frequency of those haplotypes known to present at least two of those 10 epitopes.

No.	Ep. 1	Ep. 2	Ep. 3	Ep. 4	Ep. 5	Ep. 6	Ep. 7	Ep. 8	Ep. 9	Ep. 10	Cov.	DCov
1	167-175	175-184	176-184	180-188	240-249	254-262	260-269	261-270	308-316	355-363	0,624	0,390
2	160-168	167-175	175-184	180-188	240-249	254-262	259-268	260-267	260-269	261-270	0,759	0,342
3	167-175	175-184	176-184	180-188	240-249	254-262	259-268	260-269	261-270	355-363	0,819	0,321
4	167-175	175-184	176-184	180-188	240-249	254-262	260-267	260-269	261-270	355-363	0,819	0,321
5	167-175	175-184	176-184	180-188	240-249	260-267	260-269	261-270	269-277	355-363	0,815	0,321
6	160-168	167-175	175-184	180-188	240-249	254-262	260-267	260-269	261-270	308-316	0,759	0,157
7	167-175	175-184	176-184	180-188	240-249	254-262	259-268	260-269	261-270	308-316	0,819	0,136
8	167-175	175-184	176-184	180-188	240-249	254-262	260-267	260-269	261-270	308-316	0,819	0,136
9	167-175	175-184	176-184	180-188	240-249	260-267	260-269	261-270	269-277	308-316	0,815	0,136
10	160-168	167-175	175-184	180-188	240-249	254-262	260-267	260-269	261-270	306-316	0,849	0,079
11	167-175	175-184	176-184	180-188	240-249	254-262	259-268	260-269	261-270	269-277	0,872	0,059

13. Sectors of Nef

The accessory protein Nef is essential for HIV-1 pathogenesis and participates in a multitude of signaling and trafficking pathways.

Two sectors were identified from the analysis of HIV-1 Nef sequences. As in the case of Gag, we downloaded clade B sequences only, keeping one sequence per patient.

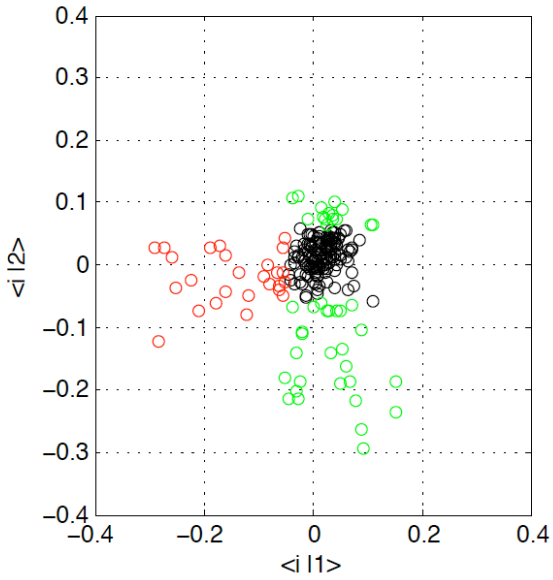

Figure S7. Projections of the components of the 2 eigenvectors corresponding to the 2 highest eigenvalues of the correlation matrix, after removal of phylogeny (see S4), for the Nef multiple sequence alignment. The projections corresponding to sector 1 sites are shown in red, and those for sector 2 sites are shown in green.

The analysis of the sequence similarity matrix also revealed phylogenetic clusters (data not shown). We kept the main cluster, comprising 1443 sequences. The eigenvector maps used to determine the sectors are shown in Fig. S7. The corresponding cleaned correlation matrix is illustrated in Fig. S8, and the sector residues are listed in Table S7.

Structural and mutational studies have characterized the structure-function relationships of Nef motifs that are known to play an important role in viral fitness[13]. A crucially important region (positions 69-79) is part of sector 1 (except position 71). This left-handed helical region is proline-rich, containing three consecutive, highly conserved PxxP motifs and is important for the interaction between Nef and the Src homology region-3 (SH3) domains of signaling molecules such as Hck, Lck and Vav [14]. While dispensable for CD4 down-regulation, this region is necessary for viral replication. Additionally, this region is also known to associate with the TCR zeta chain, which is required for the upregulation of the protein FasL. Increased expression of FasL triggers the apoptotic pathway in infected and possibly uninfected cells[15].

The organization of Nef monomers into homodimers is known to be important for its function. Using biomolecular fluorescence complementation, a recent study by Poe and Smithgall[16] identified residue positions that lined the dimerization interface. Of the six positions that were identified, five (109, 112, 115, 121 and 123) are part of sector 2 in HIV-1 Nef. Moreover, four (57, 106, 109 and 110) out of ten positions implicated in binding of Nef to CD4 are part of sector 2[17]. The other six are

not in any sector because of their high degree of conservation. This observation is consistent with reports that suggest that the integrity of dimerization interface is essential for Nef-induced CD4 downregulation.

Finally, it is also worthwhile to note that experimental evidence validates the independence of sectors 1 and sector 2 in HIV-1 Nef found by RMT. The study on Nef dimerization[16] found that mutations in the PxxP motifs (sector 1) did not disrupt dimerization. Additionally, the study reports that mutations in the dimerization interface of Nef did not disturb the three-dimensional folding required for SH3 binding activity, which is mediated by residues in sector 1.

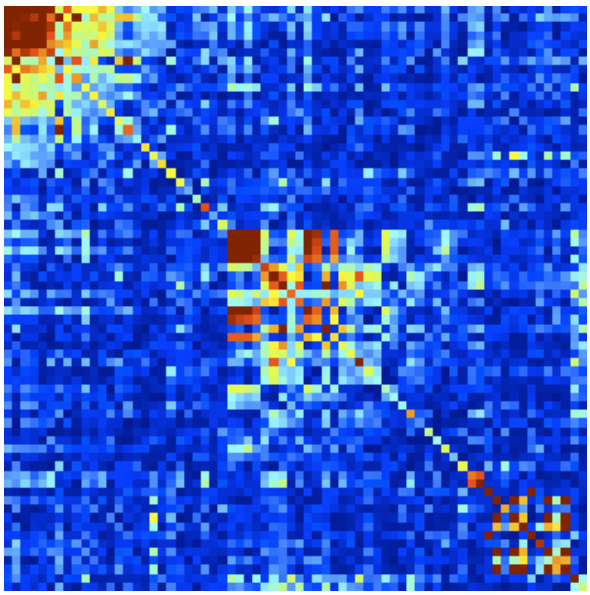

Figure S8. Absolute value of the cleaned correlation matrix for a Nef multiple sequence alignment of 1443 Clade B sequences.

14. Results for Reverse Transcriptase (RT)

We performed a similar analysis for 789 clade B RT sequences (drug naïve). There are 7 sectors. The results are reported in Figs. S9-S10 and in Table S8.

Table S7. Sites that comprise the sectors of NEF

Sector 1		Sector 2	
6	88	13	118
59	127	25	119
60	128	36	121
61	144	37	122
62	164	43	123
65		45	124
66		48	130
67		49	131
68		50	132
69		51	134
70		54	136
72		56	137
73		91	139
74		106	140
75		109	141
76		110	145
77		111	146
78		112	157
79		113	169
84		115	176
86		117	183

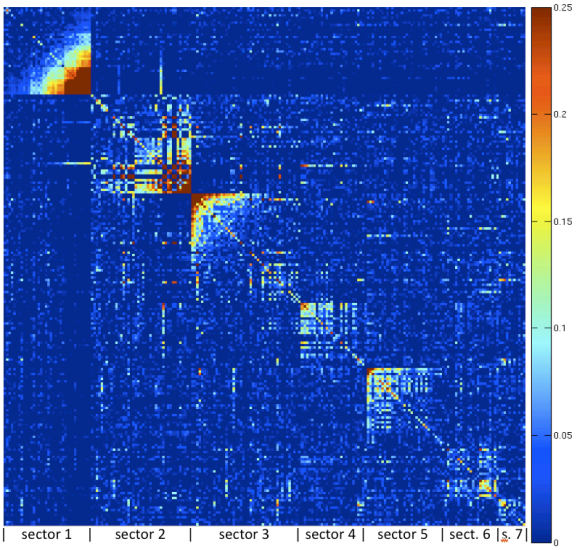

Figure S9. Absolute value of the Cleaned correlation matrix for a Reverse Transcriptase MSA of 789 Clade B sequences obtained in drug naïve individuals. The order of the rows and columns of this matrix within this representation is explained in Table S8.

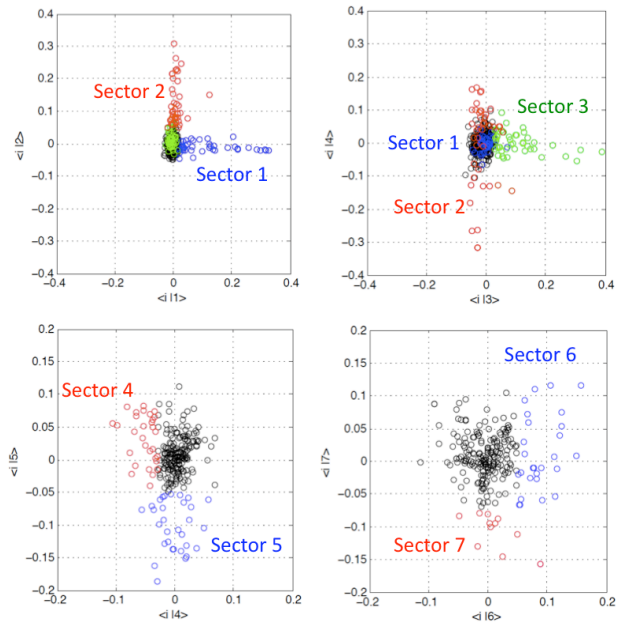

Figure S10. Projections of the components of the 7 eigenvectors corresponding to the seven highest eigenvalues of the correlation matrix, after removal of phylogeny (see S4), for the Reverse Transcriptase multiple sequence alignment.

Table S8. Sites that comprise the sectors of Reverse Transcriptase

Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6	Sector 7
15	23	9	31	44	5	81
22	40	10	53	57	6	113
80	46	12	115	125	24	124
83	47	23	160	126	35	187
106	143	34	203	134	43	258
138	158	36	222	145	60	279
156	185	41	260	147	67	308
237	190	49	264	149	69	394
238	223	50	266	153	94	395
239	230	61	270	154	103	409
240	257	62	272	164	107	417
241	268	63	278	166	118	
242	282	66	300	167	121	
243	285	70	307	168	173	
244	312	72	320	169	174	
245	335	74	330	170	184	
246	337	87	332	171	219	
247	342	89	340	175	225	
248	345	92	347	176	228	
249	346	95	349	179	234	
250	350	104	354	180	291	
251	363	114	358	181	359	
252	364	119	367	182	371	
253	365	137	374	186		
254	373	148	380	192		
255	383	162	396	227		
293	384	172	424	233		
	389	199		284		
	392	208		287		
	398	210		344		
	401	215		390		
	408	224				
	410	265				
	414	283				
	415	294				
	416	296				
	422	315				
	429	325				
	434	327				
	437	342				
	439	356				
		375				
		388				
		404				
		407				

15. Distribution of negative and positive correlations using different thresholds

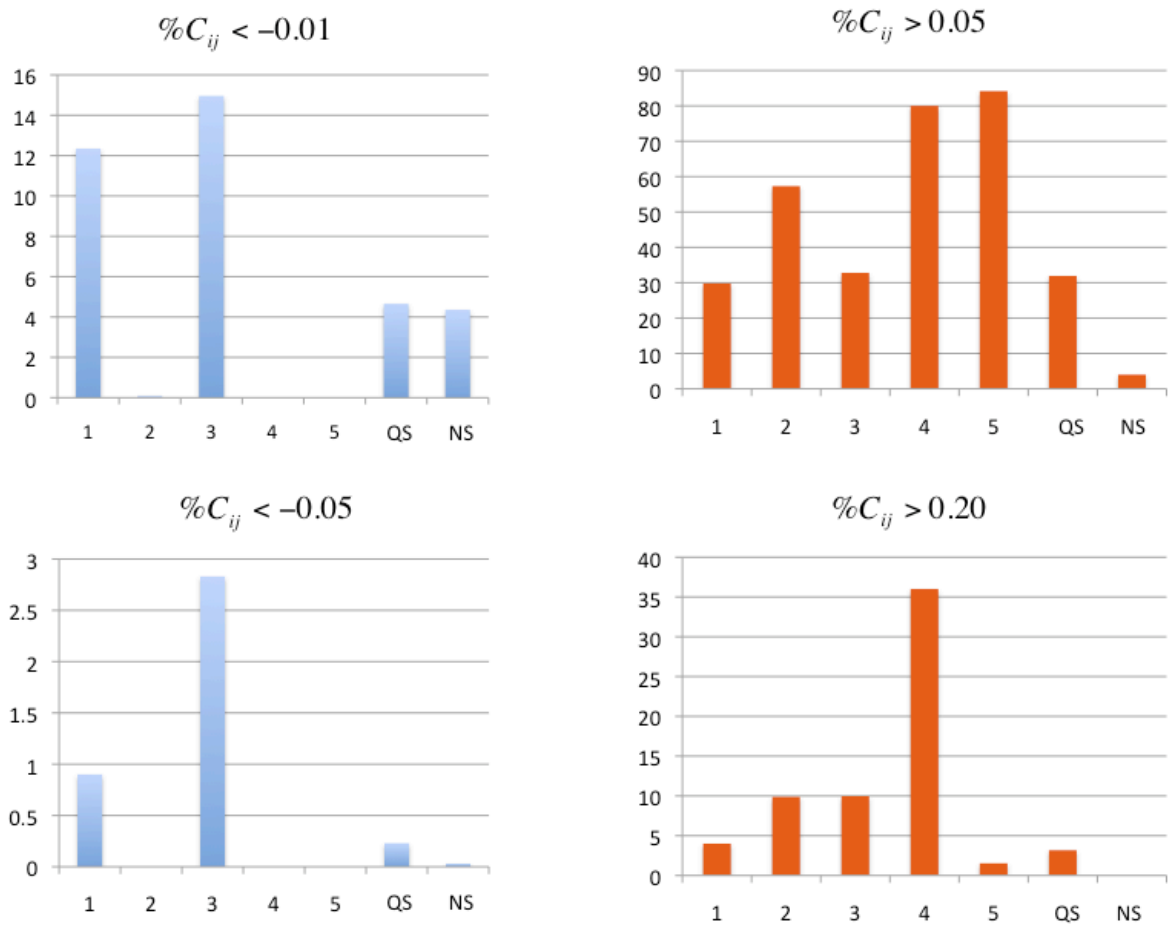

Figure S11. Percentage of significant negative or positive correlations in the different sectors using different thresholds.

16. Predictions of sign of pair correlations and comparisons with existing data

Table S9. Correlation coefficient for pair of sites for which in-vitro measurements of the fitness of single and double mutants is available. We define as compensation a case where the fitness of the double mutant is higher than the fitness of the least fit single mutant, and we define as inhibition a case where the double mutant is less fit than both single mutants.

	C_{ij}	Experiment	Reference
264-173	0.198	Compensation	Schneidwind et. al.[18]
242-219	0.152	Compensation	Brockman et. al[8]
146-147	0.057	Compensation	Troyer et. al. [19]
207-215	-0.012	Inhibition	Troyer et. al. [19]

Table S10. Prediction of pairs of sites inhibiting each other within sector 3, i.e. sites for which the double mutants are less fit than the single mutants. Those sites are all located within the p24 interfaces, inside the hexamer (Intra Hex.) or between hexamers (Hex. Hex.).

Pair	Cij (uncleaned)	Cij (cleaned)	Interface
170-295	-0.015	-0.034	Intra Hex.
171-269	-0.004	-0.032	Intra Hex.
174-295	-0.014	-0.031	Intra Hex.
186-295	-0.030	-0.030	Intra Hex.
186-269	-0.010	-0.036	Intra Hex.
198-182	-0.009	-0.036	Intra Hex.
181-310	-0.014	-0.043	Intra Hex. + Hex. Hex.
310-326	-0.045	-0.023	Hex. Hex.

Table S11. Prediction of pairs of sites inhibiting each other within sector 1, i.e. sites for which the double mutants are less fit than the single mutants.

Pair of sites	Cij (uncleaned)	Cij (cleaned)
8-153	-0.016	-0.062
11-63	-0.028	-0.057
11-86	-0.023	-0.069
12-138	-0.037	-0.052
16-141	-0.017	-0.076
24-52	-0.016	-0.068
24-87	-0.011	-0.053
24-134	-0.014	-0.059
24-160	-0.017	-0.063
38-138	-0.037	-0.095
41-144	-0.018	-0.061
63-97	-0.015	-0.081
63-149	-0.019	-0.051
94-141	-0.067	-0.063
97-138	-0.045	-0.066
99-138	-0.076	-0.063
100-148	-0.013	-0.064
148-158	-0.011	-0.056

17. Evaluation of three-site correlations

Three-site correlations reflect the way two mutations influence the probability of a third mutation. In order to evaluate 3-body correlations within each sector, we computed the correlation function between variables $x_i(s)$ and variables $x_j(s)x_k(s)$:

$$C_{ijk} = \frac{\left\langle x_i x_j x_k \right\rangle_s - \left\langle x_i \right\rangle_s \left\langle x_j x_k \right\rangle_s}{\sqrt{V_i V_{jk}}} \qquad \text{(S17)}$$

with $V_{jk} \equiv \left\langle x_j^2 x_k^2 \right\rangle_s - \left\langle x_j x_k \right\rangle_s^2$ (S18)

In Fig. 1F,G. we report the percentage of significant positive or negative 3-body correlations in each sector defined by our original RMT analyses. As before, Sectors 3 and 1 are characterized by the greater proportion of negative correlations, and few positive correlations.

18. Analysis of Sequences from Elite Controllers

Viral Gag amino acids sequences obtained from plasma of Elite Controllers infected with clade B HIV-1 were analyzed for mutations in critical sectors with respect to the Clade B consensus sequence as reference. The method for PCR and sequencing were described elsewhere before (Miura et al. JVI 82(17):8422-30). The Genbank accession numbers for the sequences previously published are EU517762 through EU517815(Miura et al. JVI 82(17):8422-30) and FJ387527(Miura et al. JVI ;83(6):2743-55); and the remainder were submitted under JF304745-JF304762.

Elite controllers were defined as HIV-1 positive individuals with plasma virus load below the limit of detection of ultrasensitive HIV RNA assays in at least three determinations that spanned a minimum of 12 months, in the absence of antiretroviral therapy. In order to keep the estimate of mutations conservative, we did not include as mutants instances where the identities of the amino acids were uncertain (marked by an 'X' or a '-' in the sequence).

Figure 3B compared the frequency of multiple mutations in sector 1 versus that in sector 3 observed in plasma sequences obtained from the viral sequences derived from 72 chronically infected patients. Figure S12 illustrates the situation when the PBMC sequences for 31 of these patients were included (n=103). As is evident, the qualitative results are similar to that of Fig. 3B. Figure S13 shows that among the few multiple mutations that we observe in sector 3, a large fraction involve residues that are positively correlated with each other.

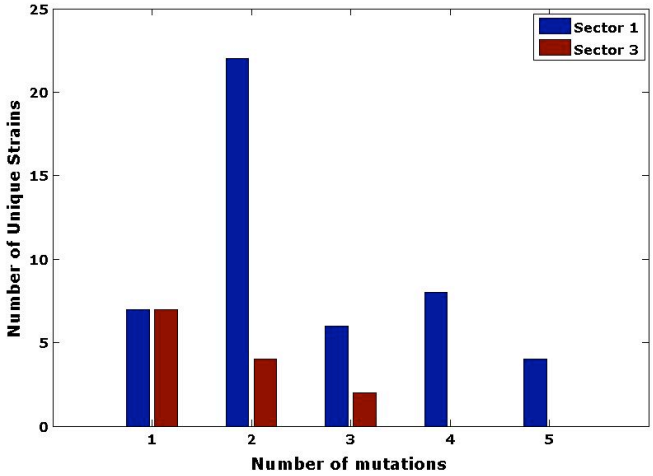

Figure S12. Comparison of the number of unique viral strains observed in plasma and PBMC sequences (n=103) derived from a cohort of elite controllers that contain different numbers of mutations in sectors 1 and 3

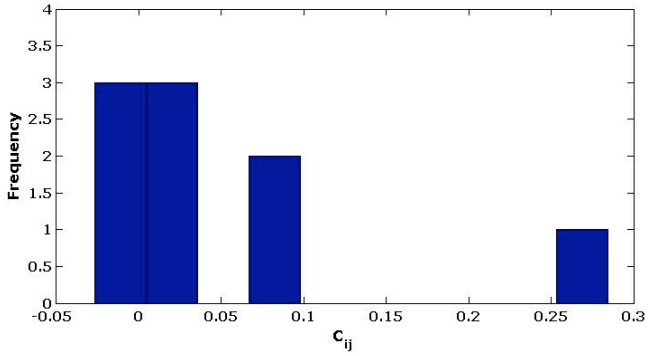

Figure S13. Comparison of the size of pairwise correlations among sector 3 residues that mutate simultaneously in the sequences derived from the cohort of elite controllers

References

1. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell*. **138**, 774-786 (2009)
2. Plerou, V. *et al.* Random matrix approach to cross correlations in financial data. *Phys Rev E Stat Nonlin Soft Matter Phys*. **65**, 066126 (2002)
3. Sakurai, J. J. a. T., S. F. *Modern Quantum Mechanics*. (Addison-Wesley Reading (Mass.), 1994).
4. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*. **315**, 1583-1586 (2007)

5. Strang, G. *Introduction to linear algebra*. (Wellesley Cambridge Pr, 2003).
6. *Los Alamos HIV Sequence Database*, <http://www.hiv.lanl.gov/>
7. Brumme, Z. L. *et al.* HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One*. **4**, e6687 (2009)
8. Brockman, M. A. *et al.* Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. *J Virol*. **81**, 12608-12618 (2007)
9. Pereyra, F. *et. al.*. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science*. **330**, 1551-1556 (2010)
10. Streeck, H. *et al.* 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 (2009)
11. Maiers, M., Gragert, L. & Klitz, W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol*. **68**, 779-788 (2007)
12. *Epitope Tables, Los Alamos HIV Molecular Immunology Database*, <http://www.hiv.lanl.gov/content/immunology/tables/tables.html>
13. Geyer, M., Fackler, O. T. & Peterlin, B. M. Structure--function relationships in HIV-1 Nef. *EMBO Rep*. **2**, 580-585 (2001)
14. Saksela, K., Cheng, G. & Baltimore, D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4. *EMBO J*. **14**, 484-491 (1995)
15. Xu, X. N. *et al.* Induction of Fas ligand expression by HIV involves the interaction of Nef with the T cell receptor zeta chain. *J Exp Med*. **189**, 1489-1496 (1999)
16. Poe, J. A. & Smithgall, T. E. HIV-1 Nef dimerization is required for Nef-mediated receptor downregulation and viral replication. *J Mol Biol*. **394**, 329-342 (2009)
17. Grzesiek, S., Stahl, S. J., Wingfield, P. T. & Bax, A. The CD4 determinant for downregulation by HIV-1 Nef directly binds to Nef. Mapping of the Nef binding surface by NMR. *Biochemistry*. **35**, 10256-10261 (1996)
18. Schneidewind, A. *et al.* Structural and functional constraints limit options for cytotoxic T-lymphocyte escape in the immunodominant HLA-B27-restricted epitope in human immunodeficiency virus type 1 capsid. *J Virol*. **82**, 5594-5605 (2008)
19. Troyer, R. M. *et al.* Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog*. **5**, e1000365 (2009)