Correlated amino acid mutation analysis has been widely used to infer functional interactions between different sites in a protein. However, this can be confounded by important phylogenetic effects broadly classifiable as background linkage disequilibrium (BLD). We have systematically separated the covariation induced by selective interactions between amino acids from background LD, using synonymous (S) vs. amino acid (A) mutations. Covariation between two amino acid mutations, (A,A), can be affected by selective interactions between amino acids, whereas covariation within (A,S) pairs or (S,S) pairs cannot. Our analysis of the pol gene — including the protease and the reverse transcriptase genes — in HIV reveals that (A,A) covariation levels are enormously higher than for either (A,S) or (S,S), and thus cannot be attributed to phylogenetic effects. The magnitude of these effects suggests that a large portion of (A,A) covariation in the HIV pol gene results from selective interactions. Inspection of the most prominent (A,A) interactions in the HIVpol gene showed that they are known sites of independently identified drug resistance mutations, and physically cluster around the drug binding site. Moreover, the specific set of (A,A) interaction pairs was reproducible in different drug treatment studies, and vanished in untreated HIV samples. The (S,S) covariation curves measured a low but detectable level of background LD in HIV.

INTRODUCTION

Correlated amino acid mutation analysis has been widely used to infer functional interactions between different sites in a protein [1–12]. Typically, a strong correlation between amino acid mutations is interpreted as evidence of functional interactions under substantial selection pressure. For example, statistical covariation of amino acid mutations in HIV has revealed interesting biological interactions between sites, and constraints imposed by protein structure [13–19]. Therefore, studying covariation of amino acid mutations in HIV will improve our understanding of HIV drug resistance as well as help vaccine design [13,18]. Studies of covariation in different regions of HIV genome have identified a number of correlated amino acid mutation pairs, many of which have known biological interactions [13–19].

However, such covariation analysis can be confounded by important phylogenetic effects [13,14]. One major challenge for covariation analysis is distinguishing covariation that is genuinely due to selection pressure, from covariation that is simply due to co-inheritance from a common ancestor. When a mutation first occurs in an individual chromosome, other mutations are already present in that chromosome, and initially this mutation will be inherited in 100% linkage with those other mutations. Such co-occurrence due to common ancestry is classified as background linkage disequilibrium (BLD) [20] (Fig. 1A). Over time, however, such linkage will be scrambled by events such as recombination and mutation, returning to equilibrium (no statistical association between them). For example, homologous recombination events between any pair of mutations will gradually scramble any linkage between the mutation pair at a rate that is proportional to the physical distance between them, the recombination rate, and the passage of time. This raises several questions. How to distinguish BLD from the covariation due to selection pressure? What fraction of covariation is BLD? How strong is BLD in HIV? The evidence from different studies has been ambiguous. On one hand, studies indicate that phylogenetic effects in HIV are strong. Phylogenetic analysis has successfully inferred HIV transmission history from HIV sequences [21,22]. On the other hand, HIV’s high mutation rate [23,24], recombination rate [25–27], and short generation time [28–30] should reduce the phylogenetic effect significantly.

It should also be emphasized that phylogenetic analysis can be confounded by strong selection pressure. Whereas phylogenetic analysis interprets the presence of the same mutation in several individuals as evidence of common ancestry, selection pressure creates bias for “convergent evolution” [31] in which the same mutation can evolve independently many times due to positive selection. Such hidden biases are incompatible with the assumptions of classical phylogenetic analysis [32]. For example, one study has reported that selection pressure for drug resistance can cause incorrect phylogenetic inferences from HIV sequences (compared with the known transmission history) [33].

Thus, it is important to develop methods that can distinguish these two causes of covariation. Recently, analytical methods such as parametric bootstrap [34] and phylogeny-based shuffling [35] have been developed to address this issue. These methods can help distinguish between phylogenetic and selective covariation, thereby improving the accuracy of functional interaction inference in HIV.

Academic Editor: Jean Carr, Institute of Human Virology, United States of America

Received July 6, 2007; Accepted August 1, 2007; Published August 29, 2007

Copyright: © 2007 Wang, Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work is supported by grants from the NIH (U54 RR021813) and DOE (DE-FG02-02ER63421), a Dreyfus Foundation Teacher-Scholar Award to CJL, as well as UCLA dissertation year fellowship and UCLA AIDS institute fellowship to QW.

Competing Interests: The authors have declared that no competing interests exist.

* To whom correspondence should be addressed. E-mail: leec@chem.ucla.edu
coevolution due to shared selection pressure. It should also be noted that selection on nucleotides (e.g., constraints on the RNA structure in viruses [36,37]) rather than on amino acids can also cause BLD, as measured by (S,S) covariation, thus may contribute to the covariation of all three types of mutation pairs, (A,A), (A,S) and (S,S), in this study.

RESULTS
Metrics of LD in HIV and Its Comparison with Background LD

First, we performed standard analyses of Linkage Disequilibrium (LD) on a dataset of about 50,000 HIV-1 pol gene sequences of subtype B, covering a 1.4 kb region of the HIV protease and reverse transcriptase (RT) genes, mostly from patients under antiretroviral drug treatment [38]. Following the procedure of the Human Genome HapMap project [39], we applied a minimum frequency criteria to the data before measuring the LD. After applying the frequency cutoff of 2%, our dataset included 398 distinct single nucleotide mutations, each with 3260 observation counts on average. It should be noted that due to the very large size of this dataset and the high rate of mutation in HIV, we detected a very high density of mutations, including mutations at the majority of individual nucleotide sites, most with large numbers of observations. This provided a uniquely high-resolution mutation dataset for mapping LD. The density of mutations (observations per nucleotide) in this dataset is 100-fold higher than in the data from the Human Genome HapMap project [39].

We computed D′ and r [40], two measures of statistical association commonly used to measure LD in many organisms, e.g., human [41–44]. Both metrics displayed a pattern in HIV similar to that in human, decaying as a function of distance (Fig. 2A, and Fig. S1A), as expected from population genetics theory. However, they indicated weaker LD than that in human [41–44], which is consistent with HIV’s high mutation rate [23,24], recombination rate [25–27], and short generation time [28–30], the factors that reduce LD according to the population genetics theory.

Another crucial difference between HIV and classical examples of LD analysis (i.e., the human data), is that this region of the HIV genome experiences very strong positive selection pressure due to antiviral drug treatments [45,46]. To control for possible LD caused by amino acid selection pressure, we repeated this LD analysis strictly for pairs of synonymous mutations. Such mutations do not change the amino acid sequence and thus are not subject to amino acid selection pressure, yielding a measurement of “background LD” free of amino acid selection artifacts [20]. D′ and r metrics of the background LD decayed as a function of physical distance (Fig. 2A, and Fig. S1A). However, the average background LD by these metrics was smaller than the average LD, across the whole one kb region. The average D′ for background LD decayed from 0.02 for adjacent mutations to 0.01 for distances of 0.4 kb or more, about two-fold lower than the standard LD curve over the same distance range (Fig. 2A). The average r showed a similar pattern (Fig. S1A). This suggests that selection pressure plays an important role in shaping LD in HIV.

Comparing Covariation of (A,A), (A,S) and (S,S) in HIV

To examine this hypothesis further, we subdivided mutation pairs into three groups: pairs of synonymous mutations (S,S); pairs of amino acid mutations (A,A); and pairs consisting of one amino acid mutation and one synonymous mutation (A,S). (A,A) pairs experience both phylogenetic effects and possible selective interactions; that is, (A,A) pairs that together increase reproductive fitness may be selected for co-occurrence. By contrast, since
Synonymous mutations do not affect the amino acid sequence, both (S,S) and (A,S) pairs are not subject to amino acid selective interaction effects. Thus, the extent of selective interaction effects can be estimated by a systematic pattern of excess covariation specifically for (A,A) pairs relative to that for (S,S) and (A,S) pairs.

We first compared the (A,A), (A,S) and (S,S) covariation in the same dataset of 50,000 HIV-1 samples. After applying minimum frequency cutoff of 2%, 124 amino acid mutations and 274 silent mutations were included, yielding 7626 (A,A) pairs, 33976 (A,S) pairs, and 37401 (S,S) pairs. Compared with (A,S) and (S,S), (A,A) covariation was dramatically higher, in both $D'$ and $r$ (Fig. 2, and Fig. S1). Only (A,A) pairs showed $D'$ greater than 0.8 and many more (A,A) pairs showed strong covariation ($D'>0.5$) than (A,S) and (S,S) pairs (Fig. 2C, 2D, 2E). Furthermore, the average covariation of (A,A) was much higher than that of (A,S) and (S,S) (Fig. 2B). The average $D'$ of (A,A) gradually declined from 0.18 to 0.03 over about 1000 bases, while the average $D'$ of (A,S) and (A,A) started at less than 0.05 and rapidly dropped to 0.01 at around 300 bases. On average, (A,A) covariation levels were two- to five-fold higher than those of (A,S) and (S,S) across this range of distances. The conclusion also held for the frequency cutoff of 1% and 4% (Fig. S2 and S3). In addition, the difference in distribution for covariation scores $D'$ of (A,A) vs. those of (A,S) and for (A,A) vs. (S,S) was statistically significant (both p-values less than $10^{-16}$, Wilcoxon rank sum test — see Materials and Methods). Thus, a predominant fraction of (A,A) covariation does not appear to be attributable to background LD as measured by (S,S) covariation.

It is also striking that the (A,S) and (S,S) covariation behaved similarly, in contrast with (A,A) covariation. The average $D'$ of (A,S) and (S,S) both started under 0.05 and gradually decayed until they reached a flat of around 0.01 at 300 bases (Fig. 2B). The same pattern was repeated in the average $r$ curve (Fig. S1B). However, it is also interesting that there appear to be slight differences between (A,S) and (S,S) at short distances (less than 200 bases). The average $D'$ value for (A,S) was significantly higher (up to 0.04) than (S,S) for adjacent mutations, but decayed more rapidly, so that this difference vanished beyond 300 bases. This higher value of (A,S) vs. (S,S) is consistent with the known strong positive selection for amino acid mutations in this region [45,46], since (A,S) pairs would be directly affected by such potential selective sweep events [47,48], whereas (S,S) pairs can only be affected indirectly (i.e. only by selective sweep for a third mutation that is a positively selected amino acid mutation).

Comparing Covariation of (A,A), (A,S) and (S,S) in the Stanford-Treated Dataset

To assess the reproducibility of these results, we repeated this analysis of (A,A), (A,S) and (S,S) covariance in a second,
imposed by antiviral drug treatment. A dominant role for selective interactions due to selection pressure to background LD as measured by (S,S) covariation, suggesting that most (A,A) covariation in this region of HIV is not attributable inherent in the dataset (Fig. 3A). These data demonstrate again largely indistinguishable within the range of sampling variance (S,S) covariation levels. Again, the (A,S) and (S,S) curves were (A,A) covariation was about two- to four-fold higher than (A,S) and (S,S), the average D$^9$ declined to 0.07 over a scale of 800 bases; while for (A,S) and (S,S), the average D$^9$ started less than 0.07 and then both fluctuated at around 0.05 (Fig. 3A). The average r showed a similar pattern, differing from D$'$ mainly in measurement scale (Fig. S4A). Overall, (A,A) covariation was about two- to four-fold higher than (A,S) and (S,S) covariation levels. Again, the (A,S) and (S,S) curves were largely indistinguishable within the range of sampling variance inherent in the dataset (Fig. 3A). These data demonstrate again that most (A,A) covariation in this region of HIV is not attributable to background LD as measured by (S,S) covariation, suggesting a dominant role for selective interactions due to selection pressure imposed by antiviral drug treatment.

To exclude the possibility that the consistent pattern between Specialty and Stanford-Treated datasets results from overlapping samples, we eliminated from the Specialty dataset all sequences with 98% or higher identity to samples in the Stanford dataset and re-analyzed the Specialty dataset. After the filtering, the (A,A) covariation level is still much higher than (A,S) and (S,S) across all distances (Fig. S5).

Comparing Covariation of (A,A), (A,S) and (S,S) in the Stanford-Un-treated Dataset

To test the role of drug-induced selection via a negative control, we carried out the same analysis in a set of samples collected from untreated patients (Stanford-Un-treated; see Materials and Methods). Previous studies have showed that comparison of these Treated vs. Untreated datasets can identify the effects of antiviral drug treatment [15,18]. The Untreated dataset contained about 4,500 drug-naive samples covering either protease or RT (see Materials and Methods). 42 amino acid mutations and 107 silent mutations (mutation frequency ≥5%; see Materials and Methods) were included in the analysis.

Strikingly, the large difference in covariation between (A,A) and (A,S)/(S,S) disappeared in the untreated dataset. The average D$'$ of (A,A) fluctuated around the average D$'$ of (A,S) and (S,S), at approximately 0.07 (Fig. 3B). The same pattern was repeated in the average r curve (Fig. S4B). These data provide a clear, independent confirmation that drug-induced selection pressure is indeed the explanation for the surplus covariation of (A,A) pairs (relative to background LD measured by (S,S)) in the Specialty and Stanford-Treated datasets, both of which included drug-treated samples.

(A,A) Pairs Near the RT Active Site Show Strong Covariation

The (A,A) covariation decay curve (Fig. 2B) revealed a clear double-peak for pairs between 400–550 base distance in the Specialty dataset. Strikingly, a similar double-peak was observed at the same location in the Stanford-Treated (A,A) curve (Fig. 3A).

We analyzed the two datasets separately to identify the mutation pairs responsible for these two peaks. These data revealed that the peaks were caused by the same set of mutation pairs in both datasets. One peak resulted from strong covariation between a cluster of mutations RT 41L and 43E with another cluster RT 67N, 70R, 210W and 219E/Q; while the other peak reflected strong covariation between a cluster RT 41L, 67N, 70R, 210W and 219E/Q with another cluster RT 208Y, 218E and 219E/Q. Interestingly, in the three-dimensional protein structure, all these residues lie close to the reverse transcriptase active site (Fig. 4), less than 25 Å apart. Furthermore, mutations RT 41L, 67N, 70R, 210W and 219E/Q are known RT drug resistance mutation [49]. Thus every single one of the (A,A) covariation pairs observed in these peaks consisted of either one, or two known drug-resistance mutations. This analysis of the individual (A,A) covariation pairs provides independent confirmation that these specific residues are positively selected for drug-resistance.

Figure 3. (A,A) Covariation Is Dramatically Higher than (A,S) and (S,S) Covariation in the Stanford-Treated Dataset but not the Stanford-Un-treated Dataset. Sliding window results of average D$'$ in A) Stanford-Treated Dataset and B) Stanford-Un-treated Dataset. Amino acid mutation pairs (A,A), red; amino acid mutations to silent mutations (A,S), blue; silent mutation pairs (S,S), green. Each sliding window contains 4% of the data points in the set.

doi:10.1371/journal.pone.0000814.g003
Comparison of Covariation Maps of Protease in the Specialty Dataset

To provide a comprehensive analysis of the specific (A,A) pairs that showed significant evidence of selective interactions, we constructed two-dimensional maps of the statistical strength of covariation between all possible pairs of codons positions in HIV protease, for (A,A), (A,S) and (S,S) (Fig. 5). We used Fisher's exact test to calculate the covariation score $h$ (see Materials and Methods), which detects statistically significant covariation even in cases with smaller counts. No minimum frequency criteria was applied. This allows us to comprehensively compare the covariation of different types of mutation pairs. In this analysis, we used the Specialty dataset, due to its much larger number of samples (about 50,000). For each pair of codon positions, the map displays the highest level of covariation measured for mutations at that pair (see Materials and Methods).

These data reveal several striking differences between (A,A), (A,S) and (S,S) covariation. First, the (A,A) map contains a large fraction (2.6%) of strong covariation effects ($h > 5$ at 95% confidence; see Materials and Methods), compared with only a small fraction for (A,S) (0.3%) and (S,S) (0.3%). Second, whereas strong (A,A) covariation broadly distributed across the whole map (Fig. 5A), in the (A,S) and (S,S) maps covariation clustered close to the diagonal (Fig. 5B and 5C), i.e. for codon positions that are close in the sequence. These data suggest that background LD decays rapidly, within 100 nt (about 30 amino acids). The fact that (A,A) covariation extended more broadly, indicates that it arises from a different process than background LD. Third, a large fraction (30 out of 53) of the codon positions identified by (A,A) covariation are known drug-resistance mutation sites [49] in HIV protease (for the list of drug-resistance codons identified, see Materials and Methods), confirming again that these covariation effects involve drug-resistance selection.

Correlation of the Covariation Between Independent Datasets

We have shown that the level of (A,A) covariation is higher than (A,S) and (S,S) covariation in both drug-treated datasets (Specialty; Stanford-Treated). However, do these independent datasets display the same covariation effects? To answer this question, we compared the covariation measurements for each (A,A) pair from these two datasets. For each (A,A) pair, we plotted the covariation value observed in the Specialty dataset against the covariation value observed in the Stanford-Treated dataset (Fig. S6). Strikingly, the (A,A) pairs that covaried in the Specialty dataset
also covaried in the Stanford-Treated (Fig. S6A), and the covariation values in the two datasets showed strong quantitative agreement, yielding a high correlation coefficient of 0.83 (See Materials and Methods) between these two independent datasets. In contrast, for (A,S) and (S,S) pairs, the covariation detected in these two datasets was low, both giving the correlation coefficients of 0.35. This indicates that a single, consistent pattern of selective interactions is reproducibly discovered in the two independent datasets, but only for (A,A) covariation.

To assess whether drug treatment acts as the consistent amino acid selection in these two datasets, we compared the (A,A) covariation in the Treated dataset with that in the Untreated one. We found that the high consistence of (A,A) covariation between the Specialty and the Treated (correlation coefficient 0.83) disappeared in the comparison between the Untreated and the Treated, leaving a correlation coefficient of only 0.39 (Fig. S6D). This suggests that drug treatment (shared by the Specialty and the Treated datasets, but not the Untreated dataset) causes the nearly identical pattern of selective interactions found in these two independent datasets.

DISCUSSION
We have systematically separated the covariation induced by selective interactions from background LD, using silent (S) and amino acid (A) mutations. Selective interactions between amino acids can be detected by (A,A) pairs, but not by (A,S) or (S,S) pairs. Our analysis of the pol gene in HIV suggests that a large portion of (A,A) covariation in HIV results from selective interactions. Meanwhile, the (S,S) covariation curves suggest a low but detectable level of background LD in HIV. Although HIV has extremely high mutation and recombination rate, as well as short generation time, the (S,S) covariation metrics were still able to detect some BLD, decreasing as a function of physical distance (Fig. 2).

Several lines of evidence demonstrate the robustness of these conclusions. First, the same results were found by three different measurements of covariation: the widely used D’ and r metrics, and Fisher’s exact test. Second, these results were reproduced in independent experimental studies (the Specialty and Stanford-Treated datasets). Third, the high level of consistency between independent (A,S) and (S,S) covariation curves suggests that the much higher level of covariation observed for (A,A) pairs cannot be attributed to background LD. Fourth, we also found direct evidence that the difference in covariation levels between (A,A) vs. (A,S)/(S,S) is due to selection, specifically, antiviral drug treatment, by comparing treated vs. untreated datasets. Fifth, the most prominent (A,A) interactions in the HIV pol gene have been independently identified as drug resistance mutations that physically cluster around the drug binding site. Finally, the specific set of (A,A) interaction pairs was reproducible in different drug treatment studies, and vanished in untreated HIV samples. Our result agrees with the ‘observation of positive epistasis in HIV’ [50].

A previous study in plastid genomes also indicates that the significant covariation in plastid genomes is likely due to changes in the selective constraints of amino acids [51].

Could the surplus of the (A,A) covariation compared with that of (A,S) and (S,S) in the treated datasets (Specialty and Stanford-Treated) be an artifact of differences in the intrinsic mutation rates between silent and amino acid mutations (e.g. silent mutations are more likely to be transitions than transversions, thus evolving faster)? We directly tested this possibility by performing the same analysis in samples from untreated patients (Stanford-Untreated). Such an artifact should have also been observed in the untreated dataset. Yet, the difference between (A,A) vs. (A,S)/(S,S) disappeared in the untreated dataset (Fig. 3), indicating that this difference was due specifically to drug-treatment. It should also be noted that in addition to drug treatment, there are other sources of selection, such as immune pressure. Like the drug-induced selection, this too only causes (A,A) but not (A,S) or (S,S) covariation. However, we didn’t detect a significant difference between (A,A) vs. (A,S)/(S,S) in the untreated samples, suggesting our approach is not sensitive enough to detect weaker selection.

How might drug treatment cause the dramatic increase in covariation of amino acid mutation pairs observed in HIV? Several models are possible. 1) Drug treatment selects for mutations that directly cause drug resistance (called primary mutations), many of which may have secondary effects such as reducing protein stability and/or other aspects of viral fitness. These mutations can in turn induce selection pressure for mutations that compensate for these effects (called accessory mutations; e.g. a mutation that restores the protein stability). Such secondary selection effects will cause a pattern of covariation of primary mutations with their associated accessory mutations. By contrast, in the untreated samples, where such positive selection forces are presumably weaker, we did not detect significant evidence of selective interactions. 2) The covariation can be caused by shared selection pressure among amino acid mutations. If mutation X and Y are independently selected for under the same drug treatment, the two mutations are likely to covary. To distinguish the aforementioned two possibilities, we need to estimate the fitness of mutation X and Y separately and compare the sum with the fitness of the XY double mutation, which is beyond the scope of this study.

Our data also indicate that accurate measurement of background LD is useful to improve the accuracy of functional interaction prediction. Accurate measurement of the background LD would enable calculation of a threshold value above which the statistics will have a specific probability of resulting from causes other than background LD. Comparison of statistical values of covariation calculated from (A,A) pairs with the background (i.e. (S,S) covariation) allows identification of pairs of sites having a specific probability of interacting due to selection on amino acids. Such phylogenetic effects should be taken into consideration in covariation analysis of amino acid interactions.

In this paper, we have only analyzed the pol gene, which is known to have experienced strong drug selection. The same method should be applied in the other regions of the HIV genome. We expect the (A,S) and (S,S) covariation, while still consistent with each other, varies across the HIV genome due to different phylogeny across the genome. The comparison of (A,A) covariation with that of (A,S)/(S,S) across the HIV genome will provide a global view of the influence of selection on mutation covariation. For example, such comparison in the env gene will hopefully improve our understanding of the interplay between the host selection and phylogeny in that region. The same analysis could also be done in subtypes other than subtype B.

MATERIALS AND METHODS
HIV-1 sequence data
The Specialty dataset contained 48,927 subtype B sequences, mostly from patients under antiretroviral drug treatment. Multiple sequence alignments and mutation detection were performed as previously described [38,52]. All these sequences covered the whole protease and part of the RT. The Treated and the Untreated datasets were downloaded from Stanford database (http://hivdb.stanford.edu/) [53], selecting only subtype B sequences. In the Treated dataset, there were 1795 protease sequences treated with protease inhibitor (this subset were used to calculate the covariation between mutations in protease) and 5121...
reverse transcriptase sequences treated with either nucleoside reverse transcriptase inhibitor (NRI) or non-nucleoside reverse transcriptase inhibitor (NNRI) (this subset were used to calculate the covariation between mutations in RT), including 1320 samples that cover both protease and reverse transcriptase sequences (this subset were used to calculate the covariation between mutations in protease and those in RT). In each subset, we required a minimum mutation frequency of 5%. In the Untreated dataset, there were 2620 PI-naive protease sequences and 1795 NRI/NNRI-naive reverse transcriptase sequences, including 1208 samples that have both protease and reverse transcriptase treatment-naive.

**Measurements of covariation for individual mutation pairs**

We used the Fisher exact test [54,55] to test for non-random associations between mutation x at position X and mutation y at position Y, by computing the p-value for the two-sided test using the 2×2 contingency table: \(N_{XaYb}\), \(N_{XbYa}\), \(N_{XaY0}\) and \(N_{X0Ya}\). \(N_{XaY0}\) is the number of samples that have mutation x at position X and but no kind of mutation at position Y; \(N_{XbY0}\) is the number of samples that have mutation x at position X and no mutation at position Y; \(N_{X0Ya}\) is the number of samples that have no mutation at position X and but no kind of mutation at position Y; \(N_{X0Y0}\) is the number of samples that have mutation at neither position. We computed the odds ratio, its confidence interval (95% two-sided) and p-value using the fisher.test function from the statistical software package R. Note: the maximum likelihood estimator for \(h\) in the Fisher's exact test is an unbiased estimator for the probability of a contingency table: \(N_{XaYb}\) vs. \(N_{X0Y0}\). Finally, we used Wilcoxon rank sum test (wilcox.test function in the R package) to compare different types of mutation pairs with respect to their covariation score (e.g., D'). The p-value is calculated for the null hypothesis that the covariation scores for the two types of mutation pairs are from the same distribution.

**Average LD as a function of distance**

Mutation pairs with negative LD were excluded. Mutation pairs were ranked by physical distance. We calculated smoothed curves using a sliding window, the window width of 2% or 4% of the total data, and an offset for neighboring windows of the 1/2 the window width.

**Covariation map**

For each pair of mutations, we used Fisher exact test to compute a p-value for statistically significant covariation, along with a lower-bound estimate for the strength of covariation \(\theta\) based on the 95% confidence interval. Only statistically significant mutation pairs (p<0.05 for a single pair, yielding a significance level of 0.01 after the Bonferroni correction) were included in our analysis. For a given codon position pair, the strongest covariation value \(\theta\) for any pair of mutations at the two positions (of the designated type: (A,A), (A,S), or (S,S)) was displayed in the map. In protease, the drug-resistant codons identified by (A,A) covariation are 10, 13, 16, 20, 24, 30, 32, 33, 35, 36, 43, 46, 47, 48, 50, 53, 54, 58, 60, 62, 63, 71, 75, 74, 82, 84, 85, 88, 90 and 93.

**Comparing covariation between datasets**

To test whether the covariation derived from two datasets, X and Y, was consistent, we plotted for every mutation pair the covariation measurement r in X as that in Y. We also calculated the correlation coefficient between the two datasets. Since correlation coefficient is very sensitive to outliers, the lowest correlation from 2000 bootstrap replicates (the R boot library) was taken as the correlation score between X and Y.

**Supporting Information**

**Figure S1** [AA] Covariation Measured by r Is Dramatically Higher than (A,S) and (S,S) Covariation in the Specialty Dataset. (A) Sliding window results of average r. All mutation pairs, black; silent mutation pairs (S,S) only, green. Each sliding window contains 4% of the data points in the set. (B) Sliding window results of average r. Amino acid mutation pairs (A,A), red; amino acid mutations to silent mutations (A,S), blue; silent mutation pairs (S,S), green. Each sliding window contains 2% of the data points in the set. (C–E) Plots of r against the physical distance (base) within the mutation pair for (C) (A,A), D) (A,S) and E) (S,S). Found at: doi:10.1371/journal.pone.0000814.s001 (0.84 MB TIF)

**Figure S2** [AA] Covariation Is Still Dramatically Higher than (A,S) and (S,S) Covariation in the Specialty Dataset Using 0.01 As the Mutation Frequency Cutoff. (A–C) Plots of D' against the physical distance (base) within the mutation pair for A) amino acid mutation pairs (A,A), B) amino acid mutations to silent mutations (A,S) and C) silent mutation pairs (S,S). (D, E) Sliding window results of average D) D' and E) r. (A,A), red; (A,S), blue; (S,S), green. Each sliding window contains 2% of the data points in the set. Found at: doi:10.1371/journal.pone.0000814.s002 (0.88 MB TIF)

**Figure S3** [AA] Covariation Is Still Dramatically Higher than (A,S) and (S,S) Covariation in the Specialty Dataset Using 0.04 As the Mutation Frequency Cutoff. (A–C) Plots of D' against the physical distance (base) within the mutation pair for A) amino acid mutation pairs (A,A), B) amino acid mutations to silent mutations (A,S) and C) silent mutation pairs (S,S). (D, E) Sliding window results of average D) D' and E) r. (A,A), red; (A,S), blue; (S,S), green. Each sliding window contains 2% of the data points in the set. Found at: doi:10.1371/journal.pone.0000814.s003 (0.55 MB TIF)

**Figure S4** [AA] Covariation Measured by r Is Dramatically Higher than (A,S) and (S,S) Covariation in the Stanford-Treated Dataset But Not the Stanford-Untreated Dataset. Sliding window results of average r in (A) Stanford-Treated Dataset and (B) Stanford-Untreated Dataset. Amino acid mutation pairs (A,A), red; amino acid mutations to silent mutations (A,S), blue; silent mutation pairs (S,S), green. Each sliding window contains 4% of the data points in the set. Found at: doi:10.1371/journal.pone.0000814.s004 (0.17 MB TIF)

**Figure S5** [AA] Covariation Is Still Dramatically Higher than (A,S) and (S,S) Covariation in the Specialty Dataset After Excluding Samples That Have Nucleotide Sequence Similarity 98% Or Greater With Any Sample In the Stanford-Treated Dataset. (A–C) Plots of D' against the physical distance (base) within the mutation pair for A) amino acid mutation pairs (A,A), B) amino acid mutations to silent mutations (A,S) and C) silent mutation pairs (S,S). (D, E) Sliding window results of average D) D' and E) r. (A,A), red; (A,S), blue; (S,S), green. Each sliding window contains 2% of the data points in the set. Found at: doi:10.1371/journal.pone.0000814.s005 (0.88 MB TIF)

**Figure S6** Shared Drug Treatment Leads To High Consistency of Amino Acid Covariation between Independent Datasets. (A–C) The covariation measurement r in the Specialty dataset plotted against that in the Stanford-Treated dataset for A) amino acid mutation pairs (A,A), B) amino acid mutations to silent mutations (A,S) and C) silent mutation pairs (S,S). (D–F) The covariation...
measurements \( r \) in the Stanford-Untreated dataset plotted against that in the Treated dataset, for D) (A,A), E) (A,S) and F) (S,S).

**Acknowledgments**

We wish to thank A. Aleksyenko, L. Chen, N. Kim, H. Lau, C. Pan and Y. Xing for helpful discussion and Prof. N. Freimer, Prof. C. Sabatti and the reviewers for their constructive comments on this manuscript.

**References**

1. Ausschütz D, Lesk AM, Bloomer AC, Klok A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 193: 693–707.
2. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 509–517.
3. Shindyalov IN, Kotechnov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7: 349–358.
4. Thomas DJ, Casari G, Sander C (1996) The prediction of protein contacts from multiple sequence alignments. Protein Eng 9: 941–945.
5. Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2: S25–32.
6. Olmea O, Rost B, Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 295: 1221–1239.
7. Larson SM, Di Nardo AA, Davidson AR (2000) Analysis of covariance in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 305: 433–446.
8. Farzelli P, Olmea O, Valencia A, Casadio R (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins Suppl 5: 157–162.
9. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56: 211–221.
10. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Struct Biol 10: 59–69.
11. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.
12. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R (2003) Allosteric determinants in guanine nucleotide-binding proteins. Proc Natl Acad Sci U S A 100: 14415–14420.
13. Thomson MM, Najera R (2005) Molecular epidemiology of HIV-1 variants in the Stanford-Untreated dataset plotted against that in the Treated dataset, for D) (A,A), E) (A,S) and F) (S,S).

**Author Contributions**

Conceived and designed the experiments: CL QW. Performed the experiments: QW. Analyzed the data: QW. Contributed reagents/materials/analysis tools: CL. Wrote the paper: CL QW.

---

**Selective Interactions vs. BLD**

---

**REFERENCES**

1. Ane C, Burleigh JG, McMahon MM, Sanderson MJ (2005) Covarion structure for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Nat Biotechnol 23: 23–35.
2. Alouf JF, Dubois LD, Samson P, Spina C, Fauchereau F, et al. (2000) A new model for HIV-1 drug resistance. Curr Opin HIV AIDS 5: 287–294.
3. Arnesen J, Mielcarek M, Sambandamurthy V, Gade S, Perelson AS, et al. (2002) Molecular footprints of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol 76: 11273–11292.
4. Arrowsmith CH, Mootz LD, Sodergren EJ, Zhang J, Betts C, et al. (2001) Molecular interaction networks in the human genome. Science 291: 92–99.
5. Arroyo DG, del Rio C, Kucharski JS, Demuth JH, Bao Y, et al. (2006) Evidence for positive epistasis in HIV-1. Science 310: 1172–1176.
6. Arvestad L, Hande KR, Elofsson A, von Heijne G (2000) Gene ontology for the whole genome. Genome Biol 1: 0029.1–0029.11.
7. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
8. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
9. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
10. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
11. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
12. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
13. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
14. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
15. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
16. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
17. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
18. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
19. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
20. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
21. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
22. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
23. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
24. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
25. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
26. Aswani SS, de la Vega P, Copeland RS, Kessner D, Blumberg RS, et al. (2002) Selective interactions vs. BLD: New views on drug resistance. PLoS Genet 8: e1000350.
52. Pan C, Kim J, Chen L, Wang Q, Lee C (2007) The HIV positive selection mutation database. Nucleic Acids Res 35: D371–375.
53. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298–303.
54. Fisher RA (1922) On the interpretation of x(2) from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85: 87–94.
55. Agresti A (1992) A survey of exact inference for contingency tables. Statistical Science 7: 131–177.
56. Brown AHD (1975) Sample Sizes Required to Detect Linkage Disequilibrium between 2 or 3 Loci. Theoretical Population Biology 8: 184–201.
57. Lewontin RC (1988) On Measures of Gametic Disequilibrium. Genetics 120: 849–852.
58. Ohta T, Kimura M (1970) Development of Associative Overdominance through Linkage Disequilibrium in Finite Populations. Genetical Research 16: 165–&.