Supporting Information

Zhang et al. 10.1073/pnas.0907304107

SI Text

Models and Computation. We modified the Bayesian epistasis association mapping (BEAM) method designed for SNP-based, genome-wide association studies to suit for the analysis of HIV drug resistance data. The original BEAM model assumes that each marker (or variable) takes on the same number of possible values, which is reasonable for SNP data. Here in drug resistance data, each variable (mutation position) can take up to 20 amino acid types. But in reality only a small number of the 20 amino acid types can be observed at each mutation position. Because the sample size of our data is fairly large, we assume that the number of possible values each variable takes is the number of distinct values observed for that variable in the data.

Details of BVP. Suppose N_t and N_u HIV protease sequences were extracted from treated ($Y = 1$) and untreated ($Y = 0$) patients, resp. These sequences have p-mutation-prone positions (we also call them "explanatory variables," or "variables," henceforth). Each position, j, can take L_j possible values. Following the description in *Methods and Materials*, we seek to partition the p-positions into three groups. Group one contains positions that independently influence the status variable Y, group two collects positions that interactively influence Y, and group zero holds those that are independent of Y. We use the indicator vector $\mathbf{I} = (I_1, ..., I_p)$ with $I_j = 0, 1, 2$ to denote the partition of the variables, and let l_0,, l_1,, and l_2 denote the numbers of variables in each group, resp.

Let $(\mathbf{X}, \mathbf{Y})$ represent all the observed data. We can write the partition of the p-positions (variables) of each observed sequence as $X_{i,G_0} = \{X_{ij} : I_j = 0\}$, $X_{i,G_1} = \{X_{ij} : I_j = 1\}$, and $X_{i,G_2} = \{X_{ij} : I_j = 2\}$, which leads to the decomposition of all the treated sequences as $\mathbf{X}^1_{G_0} = \{X_{i,G_0} : Y_i = 1\}$, $\mathbf{X}^1_{G_1} = \{X_{i,G_1} : Y_i = 1\}$, and $\mathbf{X}^1_{G_2} = \{X_{i,G_2} : Y_i = 1\}$. The corresponding decomposition of all the untreated sequences, $\mathbf{X}^0$, is $\mathbf{X}^0 = \mathbf{X}^0_{G_0} \cup \mathbf{X}^0_{G_1} \cup \mathbf{X}^0_{G_2} = \{X_{ij} : Y_i = 0\}$. Because variables in group zero are independent of Y, $\mathbf{X}^1_{G_0}$, and $\mathbf{X}^0_{G_0}$ follow the same multinomial distribution. On the other hand, observations $\mathbf{X}^1_{G_1}$ and $\mathbf{X}^1_{G_2}$ from the treated sequences on variables in groups one and two should follow distributions different from the corresponding observations $\mathbf{X}^0_{G_1}$ and $\mathbf{X}^0_{G_2}$ from the untreated sequences.

For variables in group one, we let $\Theta_1 = \{(\theta_{j1}, \theta_{j2}, ..., \theta_{jL_j})^T : I_j = 1\}$ be the mutation frequencies of the treated population, we then have

$$P(\mathbf{X}^1_{G_1}|\Theta_1, \mathbf{I}, \mathbf{Y}) = \prod_{i:Y_i=1} \prod_{j:I_j=1} P(X_{ij}|\Theta_1, \mathbf{I}, \mathbf{Y}) = \prod_{j:I_j=1} \prod_{k=1}^{L_j} (\theta_{jk})^{n_{jk}},$$

where $\{n_{j1}, ..., n_{jL_j}\}$ are the counts of different types of mutations observed at each j of all the treated sequences. Using Dirichlet (α_j) as a prior distribution for $\{\theta_{j1}, ..., \theta_{jL_j}\}$, in which $\alpha_j = (\alpha_{j1}, ..., \alpha_{jL_j})$ are called prior pseudo-counts, we integrate out $\Theta_1$to obtain that

$$P(\mathbf{X}^1_{G_1}|\mathbf{I}, \mathbf{Y}) = \prod_{j:I_j=1} \left\{ \left\{ \prod_{k=1}^{L_j} \frac{\Gamma(n_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \right\} \frac{\Gamma(|\alpha_j|)}{\Gamma(N_t + |\alpha_j|)} \right\}. \quad [S1]$$

Here, the operation $|\alpha_j|$ sums over all elements in α_j.

In group two, multiple positions contribute jointly to the drug resistance through interactions. Correspondingly, each possible mutation configuration over the l_2 variables (positions) in group two represents a potential interaction. Thus, we use a saturated multinomial distribution for $\mathbf{X}^1_{G_2}$. There are $Q = \prod_{j:I_j=2} L_j$ possible mutation configurations and each has a corresponding frequency ρ_k in the treated population. Let $\Theta_2 = (\rho_1, ..., \rho_Q)$. Then, the conditional probability of $X^1_{G_2}$ is

$$P(\mathbf{X}^1_{G_2}|\Theta_2, \mathbf{Y}, \mathbf{I}) = \prod_{k=1}^{Q} \rho_k^{n_k},$$

where n_k is the number of occurrences of mutation configuration k observed in $\mathbf{X}^1_{G_2}$. Using a Dirichlet (β) prior for Θ_2 with pseudo-counts vector $\beta = (\beta_1, ..., \beta_Q)$, we integrate out Θ_2 and obtain the marginal probability:

$$P(\mathbf{X}^1_{G_2}|\mathbf{I}, \mathbf{Y}) = \left\{ \prod_{k=1}^{Q} \frac{\Gamma(n_k + \beta_k)}{\Gamma(\beta_k)} \right\} \frac{\Gamma(|\beta|)}{\Gamma(N_t + |\beta|)}. \quad [S2]$$

The data $\mathbf{X}^1_{G_0}$ consist of observations from the treated population on positions that are not associated with the drug resistance. Hence, they follow the same distribution as $\mathbf{X}^0_{G_0}$ and can be combined with the untreated data. In the main paper, we also introduce a model indicator J_{un} so that the observations at positions in G_2 in the untreated population are modeled as mutually independent when $J_{un} = 0$, and as a saturated multinomial distribution when $J_{un} = 1$. We derive the marginal probability of $\mathbf{X}^1_{G_0} \cup \mathbf{X}^0$ under the independence model ($J_{un} = 0$) below and note that the computation is similar when $J_{un} = 1$. Let $\Theta = (\theta_1, ..., \theta_p)$denote the mutation frequencies in the untreated population over the p positions, where $\theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jL_j})$ is the frequency at position j. We have

$$P(\mathbf{X}^1_{G_0}, \mathbf{X}^0|\Theta, \mathbf{I}, \mathbf{Y}, J_{un} = 0) = \prod_{j=1}^{p} \prod_{k=1}^{L_j} \theta_{jk}^{m_{jk}},$$

where m_{jk} is the number of occurrences of mutation type k observed at position j in $\{\mathbf{X}^1_{G_0} \cup \mathbf{X}^0\}$. Using Dirichlet (αj) priors for θj with $\alpha_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jL_j})$, we integrate out the θj and obtain that

$$P(\mathbf{X}^1_{G_0}, \mathbf{X}^0|\mathbf{I}, \mathbf{Y}, J_{un} = 0) = \prod_{j=1}^{p} \left\{ \left\{ \prod_{k=1}^{L_j} \frac{\Gamma(m_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \right\} \times \frac{\Gamma(|\alpha_j|)}{\Gamma(\sum_{k=1}^{L_j} m_{jk} + |\alpha_j|)} \right\}. \quad [S3]$$

The posterior probability of the partition indicator $\mathbf{I}$ can be expressed as

$$P(\mathbf{I}|\mathbf{Y}, \mathbf{X}, J_{un} = 0) \propto P(\mathbf{X}^1_{G_1}|\mathbf{I}, \mathbf{Y}) P(\mathbf{X}^1_{G_2}|\mathbf{I}, \mathbf{Y}) \times P(\mathbf{X}^1_{G_0}, X^0|\mathbf{I}, \mathbf{Y}, J_{un} = 0) \pi(\mathbf{I}). \quad [4]$$

The prior distribution $\pi(\mathbf{I})$ of the group indicator reflects the prior knowledge of both the total number and importance of positions associated to the drug resistance. In our implementation, we chose the multinomial prior $\pi(\mathbf{I}) \propto p_1^{l_1} p_2^{l_2} (1 - p_1 - p_2)^{p - l_1 - l_2}$, in which parameters p_1, p_2 are the expected proportions of candidate positions that belong to groups one and two, resp. The prior on J_{un} was Bernoulli (0.5). We observed that the final results were insensitive to the priors on $\mathbf{I}$ and J_{un}.

Recursive Model Selection. ***Model likelihood for the chain-dependence model.*** Suppose that X_G follows a chain-dependence model and has the structure $X_A \to X_B \to X_C$. Assume that X_A takes on N_A possible values, following Multinom (Θ_A), where $\Theta_A = (\theta^A_1, ..., \theta^A_{N_A})$. The prior distribution for Θ_A is Dirichlet(β^A) with the pseudo-counts vector $\beta^A = (\beta^A_1, ..., \beta^A_{N_A})$ in which $\beta^A_k = n_0/N_A$. This prior (and others) is chosen based on the likelihood equivalent principle of the Bayesian network (1). Marginally, this prior gives n_0 pseudo-counts per variable. We ran the RMS method with $n_0 = 1, 2, 3, \ldots\ldots, 7$ and found no difference in the results (see next section). Suppose X_B takes on N_B possible values, and $P(X_B|X_A) =$ Multinom $(\Theta_{B|X_A})$ with $\Theta_{B|X_A} = (\theta^B_{1,X_A}, ..., \theta^B_{N_B,X_A})$. We let $\Theta_{B|X_A} \sim$Dirichlet $(\{\beta^{B|A}_{i|X_A}\}_{i=1,...,N_B})$ *a priori*, with $\beta^{B|A}_{i|X_A} = n_0/(N_A N_B)$. Lastly, suppose X_C takes on N_C possible values; $P(X_C|X_B) =$ Multinom$(\Theta_{C|X_B})$ with $\Theta_{C|X_B}$ following Dirichlet$(\{\beta^{C|B}_{i,X_B}\}_{i=1,...,N_C})$ *a priori*. Again, we set $\beta^{C|B}_{i,X_B} = n_0/(N_B N_C)$.

Suppose our data D consists of n-iid observations ($n = N_t$ in treated group and $n = N_u$ in untreated group) on $X_G = (X_A, X_B, X_C)$, we want to compute the likelihood of the chain-dependence model, that is , $P(D|\Pi, I_{CV} = 1)$. For simplicity, we suppress the model indication variable $I_{CV} = 1$ in the following derivation. We summarize the data as counts corresponding to X_A, $X_B|X_A$, and $X_C|X_B$, and decompose the model likelihood as

$$P(D|\Pi) = P(D_A|\Pi)P(D_B|D_A,\Pi)P(D_C|D_B,\Pi). \qquad \textbf{[S5]}$$

That is, we let $\mathbf{n}^A = (n^A_1, \ldots, n^A_{N_A})$, where n^A_k is the number observations whose X_A takes on the k^{th} configuration. Integrating out the multinomial parameters, we obtain that

$$P(D_A|\Pi) = \frac{\Gamma(\mathbf{n}^A + \beta^A)}{\Gamma(|\mathbf{n}^A| + |\beta^A|)} \frac{\Gamma(|\beta^A|)}{\Gamma(\beta^A)} \equiv \left(\prod_{k=1}^{N_A} \frac{\Gamma(n^A_k + \beta^A_k)}{\Gamma(\beta^A_k)}\right) \frac{\Gamma(\sum_{k=1}^{N_A} \beta^A_k)}{\Gamma(n + \sum_{k=1}^{N_A} \beta^A_k)}, \qquad \textbf{[S6]}$$

where we define $\Gamma(\mathbf{v}) = \Gamma(v_1)\cdots\Gamma(v_k)$ for $\mathbf{v} = (v_1, \cdots, v_k)$. Similarly, we obtain that

$$P(D_B|D_A,\Pi) = \prod_{i=1}^{N_A} \left[\frac{\Gamma(\mathbf{n}^{B|A}_{\cdot|i} + \beta^{B|A}_{\cdot|i})\Gamma(|\beta^{B|A}_{\cdot|i}|)}{\Gamma(n^A_i + |\beta^{B|A}_{\cdot|i}|)\ \Gamma(\beta^{B|A}_{\cdot|i})}\right], \qquad \textbf{[S7]}$$

where $\mathbf{n}^{B|A}_{\cdot|i} = (n^{B|A}_{1|i}, \ldots, n^{B|A}_{N_B|i})$, with $n^{B|A}_{j|i}$ recording the number of observations in which $X_A = i$ and $X_B = j$. Thus, $|\mathbf{n}^{B|A}_{\cdot|i}| = n^{B|A}_{1|i} + \cdots + n^{B|A}_{N_B|i} = n^A_i$. Finally, we get

$$P(D_C|D_B,\Pi) = \prod_{j=1}^{N_B} \left[\frac{\Gamma(\mathbf{n}^{C|B}_{\cdot|j} + \beta^{C|B}_{\cdot|j})\Gamma(|\beta^{C|B}_{\cdot|i}|)}{\Gamma(n^B_j + |\beta^{C|B}_{\cdot|j}|)\ \Gamma(\beta^{C|B}_{\cdot|j})}\right]. \qquad \textbf{[S8]}$$

Thus, the likelihood $P(D|\Pi)$ for the chain-independence model is the product of Eq. **S6** minus **S8**. If we assign a prior on Π, such as uniform, we obtain its posterior distribution by combining the prior with the likelihood $P(D|\Pi)$ and use an MCMC algorithm to sample from or optimize this posterior distribution.

Model likelihood for the V-structure model. Now suppose the set of variables X_G follow a V-structure model. Then the model likelihood (suppressing the notation $I_{CV} = 0$), can be decomposed as

$$P(D|\Pi) = P(D_A|\Pi)P(D_C|\Pi)P(D_B|D_A,D_C,\Pi). \qquad \textbf{[S9]}$$

With the same multinomial-Dirichlet distribution on X_A, we compute $P(D_A|\Pi)$ as in Eq. **S6**. Let N_C be the number of possible values X_C can take. We let $X_C \sim$ Multinom , (Θ_C), with $\Theta_C = (\theta^C_1, ..., \theta^C_{N_C})$ following the Dirichlet (β^C) distribution *a priori*, where $\beta^C = (\beta^C_1, ..., \beta^C_{N_C})$ with $\beta^C_j = n_0/N_C$. Let n^C_k be the number of times X_C takes on value k in our observations D_C. Thus, we have

$$P(D_C|\Pi) = \frac{\Gamma(\mathbf{n}^C + \beta^C)}{\Gamma(n + |\beta^C|)} \frac{\Gamma(|\beta^C|)}{\Gamma(\beta^C)}. \qquad \textbf{[S10]}$$

Finally, we let $\theta^{B|AC}_i = \{\theta^{B|AC}_{j|i}, j = 1, ..., N_B\}$ for $i = 1, ..., N_A \times N_C$ be transition probabilities between $A \cup C$ and B, and let $n^{B|AC}_{j|i}$ be the number of transitions from allele combination i of the union of A and C to allele combination j of group B. We assign a Dirichlet prior to $\theta^{B|AC}_i$ with parameter $\beta^{B|AC}_i = (\beta^{B|AC}_{j|i}, j = 1, ..., N_B)$. We set $\beta^{B|AC}_{j|i} = n_0/(N_A N_C N_B)$ in our analysis. Integrating out $\theta^{B|AC}_i$, we get

$$P(D_B|D_A,D_C,\Pi) = \prod_{i=1}^{N_A \times N_C} \left[\left(\prod_{j=1}^{N_B} \frac{\Gamma(n^{B|AC}_{j|i} + \beta^{B|AC}_{j|i})}{\Gamma(\beta^{B|AC}_{j|i})}\right) \frac{\Gamma(|\beta^{B|AC}_i|)}{\Gamma(n^{AC}_i + |\beta^{B|AC}_i|)}\right]. \qquad \textbf{[S11]}$$

The model likelihood Eq. **S9**, in this case, is thus the product of Eqs. **S6**, **S10**, and **S11**.

Robustness of BVP and RMS to prior specifications. To check the sensitivity of the posterior analysis to different prior distributions used for the group partition of the BVP model, we tested two different priors: the first prior assumes that the variables are equally likely to be unassociated, individually associated, or interactively associated with the treatment (i.e., 1/3 of all the markers to be in group zero, group one, or group two); the second prior assumes that, on average, there are only two variables in group one, two variables in group two, and all the others in group zero. For each prior, we ran 1,000 independent MCMC chains to sample from the posterior distribution. Fig. 1*B* shows the posterior probabilities for each variable to be associated with indinavir treatment interactively based on the samples from 1,000 chains. We can see that the results from using different priors are not much different from each other, thus, in this case, the posterior probabilities are not sensitive to priors. We also tested the robustness of the BVP with respect to the total number of pseudo counts: $|\alpha_j|$ in Eq. **S1**, $|\beta|$ in Eq. **S2**, and $|\alpha_j|$ in Eq. **S3**. For them being 1 and 10, we got the same results.

For the RMS method, we prescribe the prior distributions for the multinomial parameters in each candidate model according to the rules of likelihood equivalence and prior modularity described in Heckerman et al. (1995)1. Marginally, the Dirichlet prior gives n_0 pseudo-counts per variable. In other words, the "strength" of the prior belief is equivalent to n_0 observations on each of the variables. We tested the robustness of RMS with respect to the specification of n_0. For $n_0 = 1, 2, \ldots\ldots, 7$ the results for the HIV drug resistance data were the same as that shown in Fig. 2. For $n_0 = 8, 9, 10$ some ambiguities between different model types arose, but the main structures remained the same.

MCMC sampling for RMS. We use a MCMC algorithm to simulate from Eq. **1** of the main text so as to estimate the posterior distribution of I_{CV} and Π. We use the Metropolis-Hastings (MH) algorithm (2) to update the variables under the constraints that,

for the chain-dependence model, groups A and B cannot be empty, and for the V-dependence model, group A cannot be empty. Three types of proposals were used: (*i*) randomly changing a marker's group membership, (*ii*) randomly exchanging two markers between groups A, B, and C, or (*iii*) randomly switch I_{CV}. The proposed move is accepted according to the MH ratio, which is a ratio of Gamma functions. We use an annealing strategy in burn-in iterations with a temperature set high initially and gradually reduced to one.

Results for Indinavir. ***Phenotype data confirm the top interaction patterns.*** The first interaction pattern discovered by the BVP method from the treated sequences (with posterior probability >0.5) involves positions $\{24, 32, 46, 54, 82\}$. We found 24 configurations of this five-way interaction showing up at least five times in both treated and untreated samples (Table S2), of which all have significantly different frequencies in treated versus untreated samples (p-value <0.001 after Bonferroni correction). More than 96% of the untreated samples bear the wild type $\{L24, V32, M46, I54, V82\}$, while only 53.7% of the treated samples have this configuration. The occurrence frequency of configuration $\{L24, V32, M46, I54, V82I\}$ in the treated samples is also significantly lower than that in the untreated ones, which supports the hypothesis that V82I makes the virus more susceptible to indinavir (3).

For the other 22 significant configurations, they all occurred more frequently in the indinavir-treated samples than in the untreated. Phenotypic data provide confirming evidence for 21 of them: median-fold resistance (defined as fold-decreased susceptibility compared to wildtype of viruses in Stanford HIVdb website) ≥3.2, the first quartile ≥2.0, and the third quartile ≥10. The only configuration that phenotypic data do not provide confirming evidence is $\{L24, V32, M46, I54, V82T\}$, with only nine phenotypic samples. However, V82T is listed as a major indinavir resistance mutation in the Drug resistance Mutation list (4) of 2008. Because we found no positional mutation that is present in all the interaction patterns, it is possible that multiple genetic pathways and different mechanisms exist for indinavir drug resistance3.

Dependence structures of drug resistance interactions of indinavir. After finding those interacting mutations, we applied our Bayesian model selection method to detect detailed structures of interaction to those mutations: 10, 24, 32, 46, 47, 54, 71, 73, 82, and 90. Fig. 2 shows the process of RMS for inferring the dependence structure. At each step, we tried to distribute the variables into two or three subsets which entail the chain-dependence or V-dependence structure models (*Methods and Materials*). At each step if there is one structure with overwhelmingly high posterior probability (e.g., the posterior probability >0.9), we choose that structure and proceed. If there are two or more competing structures, like the case for 10 and 71, then we are not able to confidently infer the dependence structure model and in Fig. 2 we use a "?" to denote such situations.

We first applied RMS to the set of mutation positions: 10, 24, 32, 46, 47, 54, 71, 73, 82, and 90. The top structure we found has a posterior probability >0.9. This splits the 10 positions into two independent groups: group A contains 24, 32, 46, 47, 54, and 82; group C contains 10, 71, 73, and 90; and group B is empty. This special V-dependence structure suggests that the drug resistance effects of these two groups have little to do with each other (Fig. 2). Then, we applied the same method to each group and found that the group $\{10, 71, 73, 90\}$ can be further split into two smaller marginally independent groups: one of 10 and 71, the other of 73 and 90. This structure has a posterior probability >0.9. However, for the group $\{24, 32, 46, 47, 54, 82\}$, we found two competing structures each with posterior probability >0.4: one is a marginal independence structure with 24 alone in one group and all the other 5 in another, the other structure is also a marginal independence structure with 47 in one group and all the other 5 in another. This ambiguity suggests that either the data do not contain enough information for us to figure out the underlying structure, or both 24 and 47 are marginally independent (very weak interactions) with all the other four and with each other (in this case, both of the top two structures we found are correct). Similarly, we are not able to confidently detect independences (either conditional or marginal) between 10 and 71, and also between 32, 46, 54, and 82. But for 73 and 90, there is a very strong interaction between them. For 46, 54, and 82 (which are known as the main mutations for indinavir drug resistance), a very strong chain-dependence structure is detected with a posterior probability almost one. That is, 46 and 54 are conditionally independent given 82, denoted as $42 \perp 54 | 82$ in Fig. 2.

Double mutations at positions of 46 and 54 rescue the resistance caused by mutation at position 46. From Table 1 we found that, although a single mutation at 46 decreases the binding affinity of indinavir, 46 and 54 together do not. To understand how mutations at positions 46 and 54 compensate each other, we conducted free energy decomposition analyses on M46I and M46I/I54V, as shown in Fig. 3. There are 14 residues with an absolute value of $\Delta\Delta G$ larger than 0.75 kcal/mol (Fig. 3*A1*). Nine of these residues form more favorable interactions with indinavir in the M46I/I54V mutant than in the M46I mutant, and the remaining five residues do the opposite. Fig. 3*B1* shows the structural distributions of these important residues in the HIV-1 protease. Compared with the single mutation M46I, the additional mutation at 54 also significantly changes the conformation of the binding pocket (The p-value associated with the difference in the mean RMSD values of ligand in M46I single mutant versus that in M46I/I54V double mutant is $<10^{-20}$), which affects the interactions between indinavir and residues located distantly from the mutated positions of 46 and 54, and neutralizing the loss of affinity caused by the mutation at position 46. Such a conformational change of the binding pocket can be seen clearly when the average structures of the M46I/I54V and the M46I complexes are superimposed (Fig. 3*C1*).

Double mutations at positions of 54 and 82 amplify the resistance caused by mutations at position 82. Although the single I54V mutation does not weaken the binding of indinavir, it reinforces the virus' drug resistance capability caused by the V82A mutation. Fig. 3*A1–C1* shows the comparison between free energy decomposition results of the I54V/V82A double mutations and the V82A single mutation. There are 14 residues with an absolute value of $\Delta\Delta G$ larger than 0.75 kcal/mol (Fig. 3*A2*), among which seven form more favorable interactions with indinavir in the I54V/V82A mutant than in the V82A mutant. Fig. 3*B2* shows the structural distributions of these important residues in the HIV-1 protease. Similar to the M46I/I54V double mutations, the mutation at position 54 also significantly changes the conformation of the binding pocket (The p-value associated with the difference in the mean RMSD values of ligand in V82A single mutant versus that in I54V/V82A double mutant is $9.85 * 10^{-19}$), as shown by the alignment of the average structures of the I54V/V82A and the V82A complexes in Fig. 3*C2* that affects the interactions between indinavir and residues located distantly from the mutated positions of 54 and 82 and may have amplified the drug resistance caused by the mutation at position 82.

Group of 73 and 90: Another mechanism. For the three mutations in group two (i.e., G73S, L90M, and G73S/L90M), the calculated relative binding free energies are 0.85, -0.37, and 0.51 kcal/mol (Table 1), resp., suggesting that these three mutations do not

significantly impair the binding of indinavir. Because G73S usually occurs with L90M and the mechanism by which L90M causes PI resistance is still unknown (3). These calculations are consistent with observations made in previous experiments (2, 5) In Liu et al., (5) little change was observed in the inhibitory effect of indinavir to the G73S mutant ($k_i = 0.55$ nM) compared with the wild-type protease ($k_i = 0.54$ nM), while obvious differences were observed for inhibition by different substrates. In Mahalingam et al. (2) the non-active site mutant, L90M, was observed to slightly improve the inhibition by indinavir but lower the dimer's stability. Taking together our calculations and previous experiments, we hypothesize that the non-active site mutations G73S and L90M may cause drug resistance by affecting the inhibition of substrates or the dimer stability of the HIV-1 protease instead of impairing the binding of indinavir. The facts that the binding free energy analysis can explain very well the resistance of the $\{46, 54, 82\}$ group but not the $\{73, 90\}$ group, and that they form independent interaction groups statistically indicate that these two groups may cause drug resistance through two independent mechanisms.

Results for zidovudine. Zidovudine is not designed to bind to RT and block the function of RT (unlike indinavir and nevirapine in the following), but rather to compete with natural dNTPs for incorporation into the newly synthesized DNA chains where it causes chain termination. Therefore, we cannot investigate its structural basis of resistant mutations using MD simulations and free energy decompositions like what we do for indinavir and nevirapine. To date, three biochemical mechanisms of NRTI drug resistance have been identified or proposedS4[22]. The first mechanism enables RT to discriminate against NRTIs from the analogous dNTP, so the NRTIs are not added to the growing DNA chain. The second mechanism alters RT-template/primer interactions, and thus may influence subsequent NRTI incorporation. The third mechanism increases the rate of removal of the chain-terminating NRTI residue from the 3' end of the primer and enables DNA synthesis. These different resistance mechanisms seem to correlate with different sets of mutations in RT[22]. But further biochemical investigations are needed to confirm which mechanism corresponds to which independent mutation set.

Results for nevirapine.

Molecular dynamic simulation results.
As the case of indinavir, we also conducted molecular dynamics simulations and free energy calculations for single mutations we found for nevirapine. The predicted binding free energies and the corresponding energy components for the wild-type and five mutated RT/nevirapine complexes are shown in Table 2. Among all energy components, the van der Waals interaction ($-39.4 \sim -42.6$ kcal/mol) is much more favorable for the nevirapine's binding than the other energy terms, whereas the total electrostatic contribution to inhibitor binding ($\Delta Eele + \Delta GGB$) is unfavorable ($12.9 \sim 15.2$ kcal/mol). The correlations between the predicted binding free energies and each of the three important energy terms, $\Delta Evdw$, $\Delta Eele$, and ΔGPB, were calculated, and the corresponding correlation coefficients are 0.84, 0.63, and -0.29, resp., implying that the drug resistant mutations primarily affect the van der Waals interactions between nevirapine and RT.

The mechanism of drug resistance caused by single mutations.
According to the experimental data, the five mutation patterns we studied here are the most common mutations in viruses from patients with virologic failure (6). According to Table 2, for the five mutations we studied here, three of them lead to the obvious decrease of the total binding free energy of nevirapine, including G190A (3.02 kcal/mo), K103N (0.72 kcal/mol), and Y188C (3.33 kcal/mol), and four of them lead to the decrease of the van der Waals interactions between nevirapine and RT, including G190A (2.23 kcal/mol), V106N (0.59 kcal/mol), Y181C (0.37 kcal/mol), and Y188C (2.48 kcal/mol).

In order to make a thorough investigation of influence of the mutations to the binding of nevirapine, the nevirapine/residue interactions in each of the mutated complexes and the wild-type complex were decomposed and compared to one another. Fig. S3 shows the subtraction between the nevirapine/residue interactions of the wild-type complex and those of the mutated complexes (the residues with absolute difference larger than 1.5 kcal/mol were labeled).

Drug resistance caused by K103N. In Fig. S3, for K103N no residue was found to have large difference of binding free energy; that is to say, the K103N mutation does not significantly change the binding mode of nevirapine in the active site of RT. This is consistent with the previous structural investigation of mutations at 103. Structural Studies of HIV-1 RT with K103N, in both unliganded and bound to an NNRTI, have shown that the structure is only minimally changed in that, in the unliganded form, it forms a network of hydrogen bonds that are not present in the wild-type enzyme[24]. It is likely that those changes stabilize the closed pocket form of the enzyme and interfere with the ability of inhibitors to bind to the enzyme.

Drug resistance caused by other single mutations.
For the other four mutation patterns, $2 \sim 6$ residues were found to have large differences of binding free energies. According to Fig. S6, the interactions between nevirapine and the mutated residue usually decrease significantly. For example, at position 106, the loss of interaction caused by the V106A mutation is ~3.0 kcal/mol. At position 181, the loss of the interaction caused by the Y181C mutation is ~4.0 kcal/mol. At position 188, the loss of the interaction caused by the Y188C mutation is ~4.0 kcal/mol. Therefore, the loss of the binding of the mutated residue is an important contributor for the loss of the binding free energies of nevirapine. Another observation from Fig. 1 is that some residues, in addition, to the mutated residues are also involved in the change of the binding. For example, the V106A mutation impair the interactions between nevirapine and two residues at positions 181 and 188 while enhance the interactions between nevirapine and three residues at positions 227, 235, and 236.

Clinical implications of dependence structures of drug resistance interactions. For clinical practitioners, not only a drug resistance mutation list, like the one published by the IAS–USA Drug Resistance Mutations Group (Fig. 1*A*), is very useful, but also a table that provides information about interaction patterns and the probability that the HIV virus in the infected patient's body will resist to the drug given the configuration of these interaction patterns (a table, like Table S2, with a probability or risk attached to each configuration) will be very handy for clinicians treating these patients. Knowing the conditional or marginal independence structures of these interaction patterns will greatly reduce the size of such tables. For example, since we know group $\{24, 32, 46, 47, 54, 82\}$ and group $\{73, 90\}$ are marginally independent, it is not necessary to make a large table containing all the markers of 24, 32, 46, 47, 54, 73, 82, and 90. Instead, we only need two smaller tables, one for each group.

Let A denote the configurations of group $\{24, 32, 46, 47, 54, 82\}$. Suppose there are totallyG_A configurations, then A can take values from 1–G_A. Let B denote the configurations of group 73 and 90, taking values from 1–G_B. Let R denote whether the virus resists to the drug: $R = 1$ or $R = 0$. The odds (the ratio of chance that a virus resists to a drug to the chance that

it does not) of a configuration of AB is $w_{AB}=\frac{P(R=1|AB)}{P(R=0|AB)}=\frac{P(R=1)}{P(R=0)}\frac{P(AB|R=1)}{P(AB|R=0)}$.

Because group $\{24,32,46,47,54,82\}$ and group $\{73,90\}$ are marginally independent in both populations ($R=1$ and $R=0$), the odds can be written as $w_{AB}=\frac{P(R=1|AB)}{P(R=0|AB)}=\frac{P(R=1)}{P(R=0)}\frac{P(AB|R=1)}{P(AB|R=0)}=\frac{P(R=1)}{P(R=0)}\frac{P(A|R=1)}{P(A|R=0)}\frac{P(B|R=1)}{P(B|R=0)}$. The odds ratio of a mutant type AB to wildtype A_0B_0 is

$$\frac{w_{AB}}{w_{A_0B_0}}=\frac{\frac{P(R=1)P(A|R=1)P(B|R=1)}{P(R=0)P(A|R=0)P(B|R=0)}}{\frac{P(R=1)P(A_0|R=1)P(B_0|R=1)}{P(R=0)P(A_0|R=0)P(B_0|R=0)}}=\frac{\frac{P(R=1|A)P(R=1|B)}{P(R=0|A)P(R=0|B)}}{\frac{P(R=1|A_0)P(R=1|B_0)}{P(R=0|A_0)P(R=0|B_0)}}=\frac{w_Aw_B}{w_{A_0}w_{B_0}},$$

, which is equivalent to $\log\frac{w_{AB}}{w_{A_0B_0}}=\log\frac{w_A}{w_{A_0}}+\log\frac{w_B}{w_{B_0}}$.

Thus, if we have tabulated the logarithm of the odds ratios for each configuration in each group, the log-odds ratios for the configurations resulting from the combination of the two groups of variables is trivial to derive. Similarly, if groups A and C are conditionally independent given the group B, just as 46 and 54 are conditionally independent given 82, we have

$$\log\frac{w_{ABC}}{w_{A_0B_0C_0}}=\log\frac{w_B}{w_{B_0}}+\log\frac{w_{AB}}{w_{A_0B_0}}+\log\frac{w_{BC}}{w_{B_0C_0}}.$$

Instead of a large table of all the configurations in the three groups, we need three smaller tables: one for group B, one for the union of group A and B, and one for the union of group B and C. The log-odds ratio of a configuration of A, B, and C is just the summation of the corresponding terms in these three tables. In this way, clinicians are able to tell, given the genotype of the virus from an infected patient, what is the relative risk of resistance to a drug compared to a patient with a wild-type isolate.

Molecular Mechanics (MM) Minimizations and Molecular Dynamics (MD) Simulations. ***Preparation of the initial structures.*** We have studied indinavir binding to the wild-type HIV-1 protease and its ten mutants. The crystal structure of the HIV-1 protease complexed with indinavir [(PDB entry: 1hsg (7)] was used as the starting structure. All missing hydrogen atoms of the protein were added using the tleap program in AMBER9.0. (3) The side chains of the wild-type protease were then mutated using the scap program (8) to model the complex structures of resistant mutants. We considered five single mutations, M46I, I54V, L90M, G73S, and V82A; four double mutations, M46I/I54V, M46I/V82A, 54V/82A, and G73S/L90M; and one triple mutation, M46I/I54V/V82A. These mutations can be divided into two groups: group one involves two positions of 73 and 90, and group two involves residues mutated at positions 46, 54, and 82. Moreover, we have studied nevirapine binding to the wild-type HIV-1 reverse transcriptase and its five mutants. The crystal structure of the HIV-1 reverse transcriptase complexed with nevirapine (PDB entry: 3hvt) (4) was used as the starting structure. We considered five single mutations, including G190A, K103N, V106A, Y181C, and Y188C.

Indinavir and nevirapine were optimized by the semiempirical AM1 method, and the atom partial charges were derived by fitting the electrostatic potentials to the single-point Hartree-Fock (HF)/6-31G* calculations using the RESP fitting technique (9). The quantum optimization and the electrostatic potentials were calculated using Gaussian 98 (10). In the MM minimizations and MD simulations, AMBER03 force field was used for proteins (11) and general AMBER force field (gaff) (6) was used for drugs. Partial charges and force field parameters for indinavir and nevirapine were automatically assigned using the antechamber program in AMBER (12). Each complex was immersed in a rectangular box of TIP3P water molecules. The water box extended 10Å away from any solute atom.

Molecular dynamics (MD) simulations. In MM minimization and MD simulations, particle mesh Ewald (PME) was employed for considering long-range electrostatic interactions (13). Prior to the MD simulations, the systems were optimized by 1,000 steps of steepest descent minimizations by constraining all C_α atoms with a harmonic force of 50 kcal/mol/ Å^2 and the following 1,000 steps of steepest descent minimizations by constraining all C_α atoms with a harmonic force of 10 kcal/mol/ Å^2. Subsequently, MM minimization (1,000 steps of steepest descent and 4,000 steps of conjugate gradient minimization) was used to optimize the structures without any constraint. Next, the system was gradually heated from 10 to 300 K over 25 ps using NVT ensemble. Initial velocities were assigned from a Maxwellian distribution at the starting temperature. We conducted 3 and 2 ns MD simulations for the protease and the reverse transcriptase complexes, resp., at the NPT ensemble with a target temperature of 300 K and a target pressure of 1 atm. The SHAKE procedure was employed to constrain all hydrogen atoms (14), and the time step was set to 2.0 fs. During the sampling process, coordinates were saved every 4 ps. The MM optimization and MD simulations were accomplished using the sander program in AMBER9.0.

MM/GBSA calculations. The absolute binding free energy (ΔG_{bind}) was calculated by using the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) technique according to Eq. **S12**; (5, 15, 16)

$$\begin{aligned}\Delta G_{bind}&=\Delta H-T\Delta S\approx\Delta E_{MM}+\Delta G_{sol}-T\Delta S\\&=\Delta E_{ele}+\Delta E_{vdw}+\Delta G_{GB}+\Delta G_{SA}-T\Delta S.\end{aligned}\qquad\textbf{[S12]}$$

The gas-phase interaction energy between protease and ligand, ΔE_{MM}, denotes the sum of MM energies of the molecules from electrostatic (ΔE_{ele}) and van der Waals (ΔE_{vdw}) energies. The solvation free energy $\Delta G_{solvation}$ is computed as the sum of polar (ΔG_{GB}) and non-polar (ΔG_{SA}) contributions. The ΔG_{GB} was calculated by using the modified generalized born (GB) model ($igb=2$) developed by Onufriev and coworkers (2). The exterior and the interior dielectric constants were set to 80 and 2, resp. The nonpolar contribution was computed based on solvent-accessible surface area (SASA) using the LCPO method (17): $\Delta G_{SA}=0.0072\times\Delta SASA.$ $-T\Delta S$ is the conformational entropy change upon binding, which was not considered in this study due to high computational cost (15). Each energy term was calculated using 125 snapshots extracted from the MD simulations between 0.5 and 3.0 ns for the protease complexes and 75 snapshots extracted from the MD simulations between 0.5 ns and 2.0 ns for the reverse transcriptase complexes.

Inhibitor-residue free energy decomposition analysis. The interaction between indinavir and the protease was decomposed by using the MM/GBSA decomposition procedure implemented in AMBER9 (3). We calculated van der Waals (ΔE_{vdw}), electrostatic (ΔE_{ele}), and desolvation ($\Delta G_{GB}+\Delta G_{SA}$) energies between the inhibitor and each of the protease residue (Eq. **S13**).

$$\Delta G_{inhibitor-residue}=\Delta E_{vdw}+\Delta E_{ele}+\Delta G_{GB}+\Delta G_{SA}\qquad\textbf{[S13]}$$

The polar contribution (ΔG_{GB}) of desolvation was computed using the modified GB model developed by Onufriev and coworkers (2). The non-polar contribution of desolvation (ΔG_{SA}) was computed using the surface area. The charges used in GB calculations were taken from the AMBER parameter set. All energy components were calculated using 125 snapshots from 0.5 ns to 3.0 ns for the protease complexes and 75 snapshots from 0.5 ns to 2.0 ns for the reverse transcriptase complexes.

Results and discussions of MD simulations for the protease complexes. Root-mean-square deviations (RMSDs) of the protease C_α atoms in the MD simulation relative to the starting structures were calculated (Fig. S2). The RMSD plots indicate that these complexes reached equilibrium at about 500 ps.

Free-energy decomposition analyses showed that the most favorable energy component to inhibitor binding is van der Waals ($-70.8 \sim -83.3$ kcal/mol) (Table 1). The favorable electrostatic energies ΔE_{ele} are partially cancelled by the unfavorable electrostatic solvation energies ΔG_{GB}. The total electrostatic contribution to inhibitor binding ($\Delta E_{ele} + \Delta G_{GB}$) is slightly unfavorable ($8.4 \sim 12.2$ kcal/mol). Non-polar solvation terms, which correspond to the burial of solvent accessible surface area (SASA) upon binding, contribute slightly favorably ($-9.6 \sim -10.4$ kcal/mol). We also computed the correlations between the measured binding free energies and each of the four free energy components, ΔE_{vdw}, ΔE_{ele}, ΔG_{PB}, and ΔG_{SA}, and the corresponding correlation coefficients are 0.94, 0.42, −0.43, and 0.08, resp. It is obvious that the drug resistant mutations primarily affect the van der Waals interactions between indinavir and the protease.

1. Heckerman D, Geiger D, and Chickering D (1995) Learning Bayesian networks: The combination of knowledge and statistical data, *Mach Learn* 20:197–243.
2. Onufriev A, Bashford D, and Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55:383–394.
3. Case D, et al. (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26:1668–1688.
4. Smerdon SJ, et al. (1994) Structure of the binding-site for nonnucleoside inhibitors of the reverse-transcriptase of human-immunodeficiency-virus type-1. *Proc Nat Acad Sci USA* 91:3911–3915.
5. Kollman PA, et al. (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts Chem Res* 33:889–897.
6. Wang JM, Wolf RM, Caldwell JW, Kollman PA, and Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25:1157–1174.
7. Chen ZG, et al. (1994) Crystal-Structure at 1.9-angstrom resolution of human-immunodeficiency-virus (Hiv)-Ii protease complexed with L-735,524, an orally bioavailable inhibitor of the Hiv proteases. *J Biol Chem* 269:26344–26348.
8. Xiang ZX and Honig B (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Molec Biol* 311:421–430.
9. Bayly CI, CieplakP, Cornell WD, and Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges—The resp model. *J Phys Chem* 97:10269-10280.
10. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, et al. Gaussian 98, Gaussian Inc.; Pittsburgh, PA. 1998.
11. Duan Y et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24:1999–2012.
12. Wang JM, Wang W, Kollman PA, and Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Molec Graph Model* 25:247–260.
13. Darden T, York D, and Pedersen L (1993) Particle mesh Ewald—AN N.Log(N) method for Ewald sums in large systems. *J Chem Phys* 98:10089–10092.
14. Ryckaert JP, Ciccotti G, and Berendsen HJC (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Chem* 23:327–341.
15. Wang JM, Hou TJ, and Xu XJ (2006) Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Curr Comp-Aid Drug* 2:287–306.
16. Wang W, Donini O, Reyes CM, and Kollman PA. (2001) Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Ann Rev Biophys Biomolec Struct* 30:211–243.
17. Weiser J, Shenkin PS, and Still WC (1000) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J Comput Chem* 20:217–230.

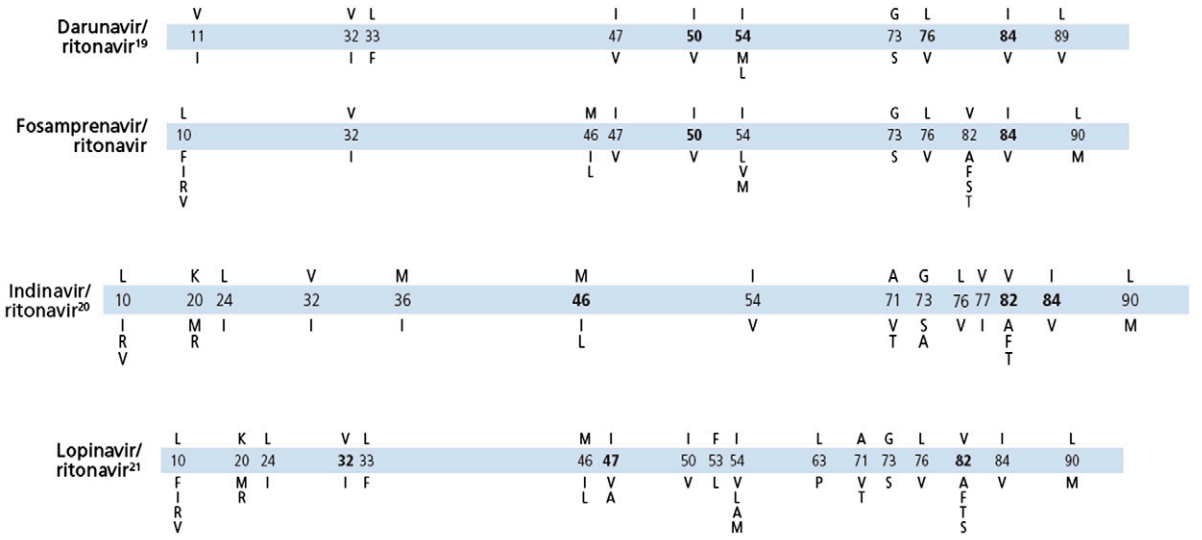

Fig. S1. The snapshot from drug resistance mutation list (6), updated spring 2008, and the posterior probabilities of each mutation position to be associated interactively with indinavir treatment. *A*) A snapshot from drug resistance mutation list (4) update spring 2008 [The International AIDS Society—USA publishes a drug resistant mutation list for every Food and Drug Administration-approved drug and updates it twice each year (4)]. The one highlighted in *Circle* is the drug we investigated in this paper, indinavir.

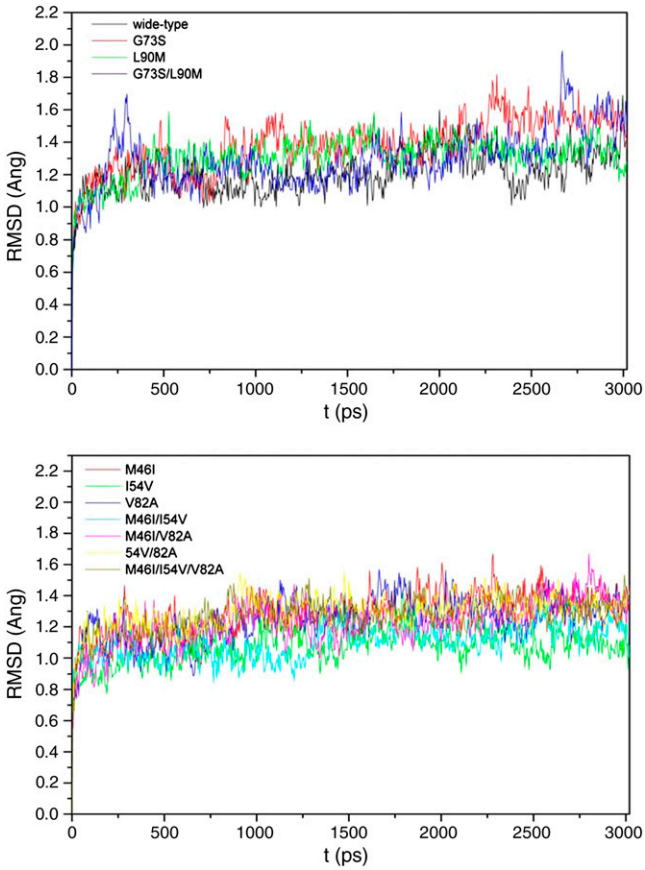

Fig. S2. Root-mean-square deviation (RMSD) of the backbone C_α atoms of (*a*) the wild-type HIV-1 protease complex and three mutated complexes in group 1 and (*b*) seven mutated complexes in group 2.

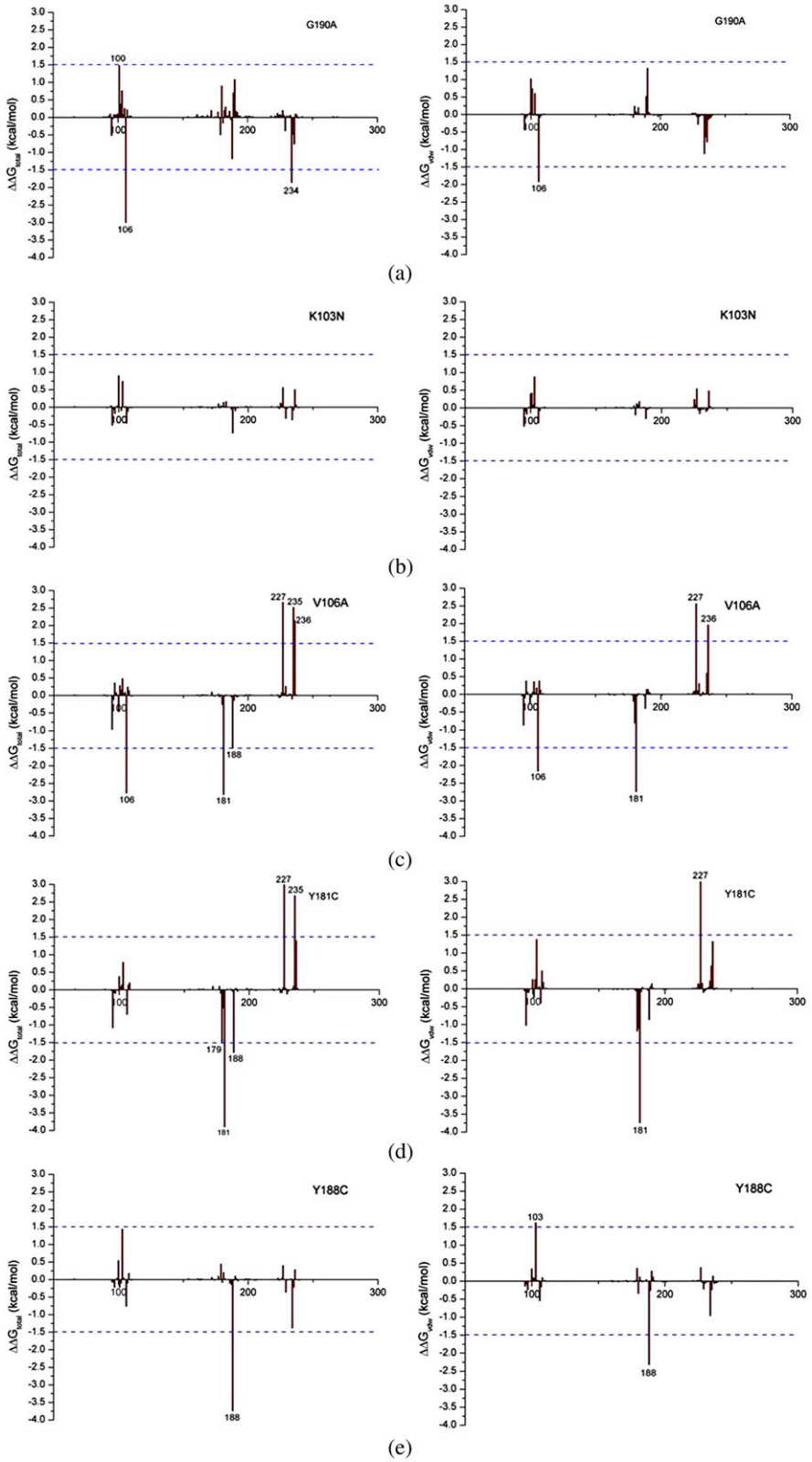

Fig. S3. The difference between the nevirapine/RT residue interactions of the wild-type complex and those of the (*a*) G190A, (*b*) K103N, (*c*) V106A, (*d*) Y181C, and (*e*) Y188C mutated complexes. Left is the total binding free energy and right is the van der Waals interactions. The residues with absolute value larger than 1.50 kcal/mol are labeled. (ΔΔG equals to the interactions between nevirapine and the wild-type WT minus those between nevirapine and the mutated RT).

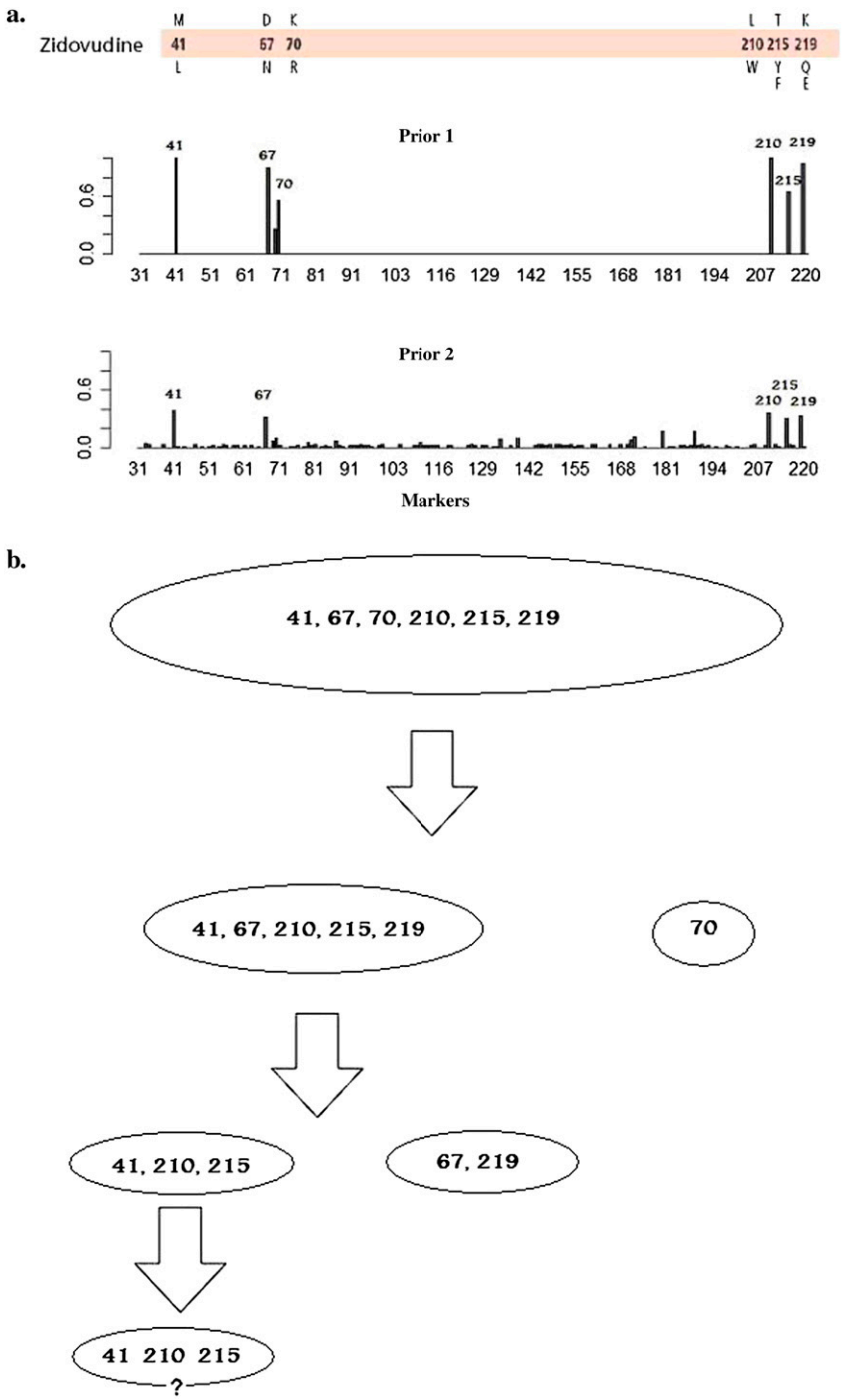

Fig. S4. (*a*) Zidovudine-resistant mutations from Drug resistance Mutation list updated in spring 2008 and the posterior probabilities of each marker to be associated interactively with zidovudine treatment. (*b*) The procedure and results of detecting detail structure of interaction of resistance to zidovudine.

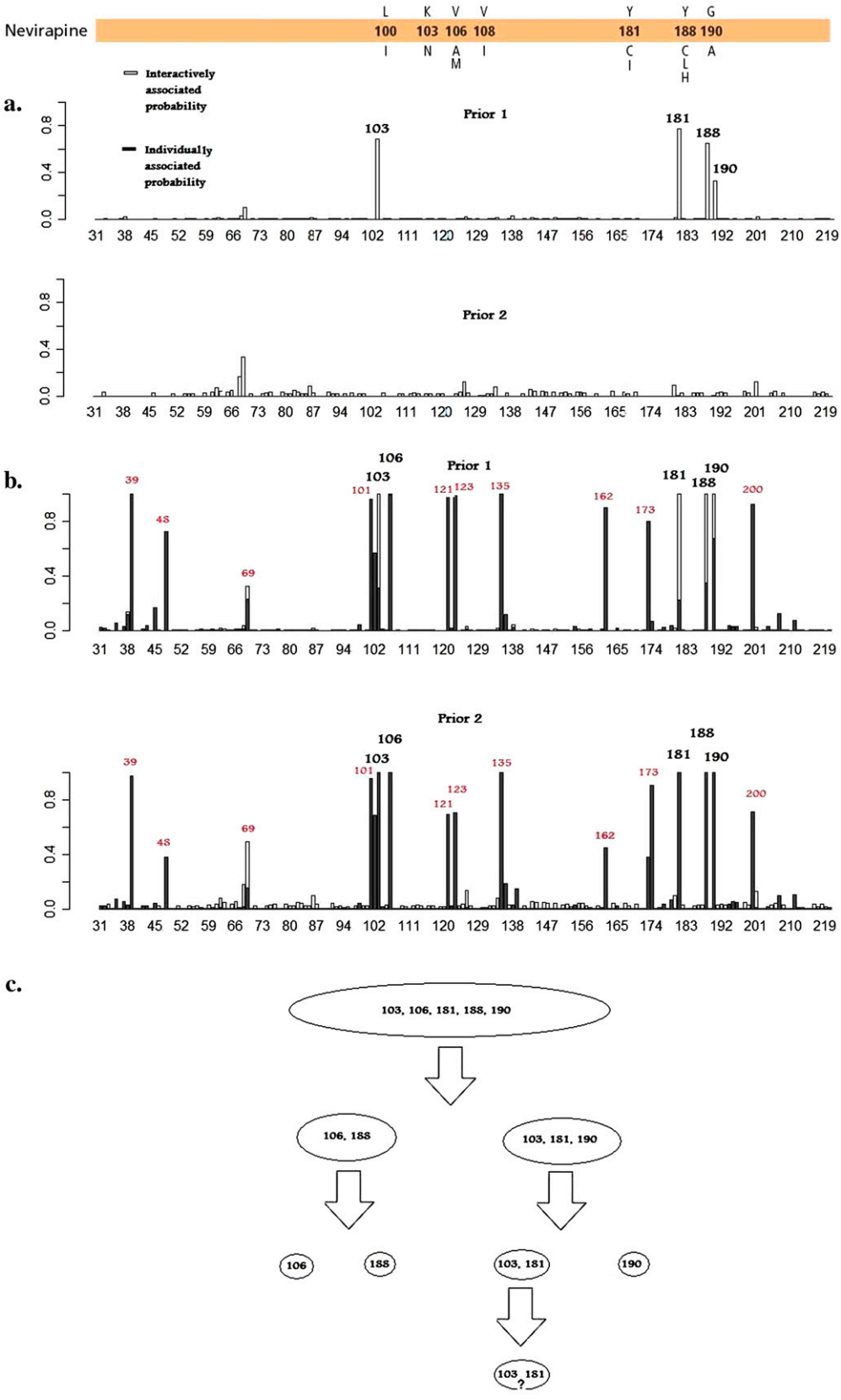

Fig. S5. (*a*) Nevirapine-resistant mutations from Drug resistance Mutation list updated in spring 2008 and the posterior probabilities of each marker to be associated (both individually and interactively) with nevirapine treatment. (*b*) The procedure and results of detecting detail structure of interaction of resistance to nevirapine.

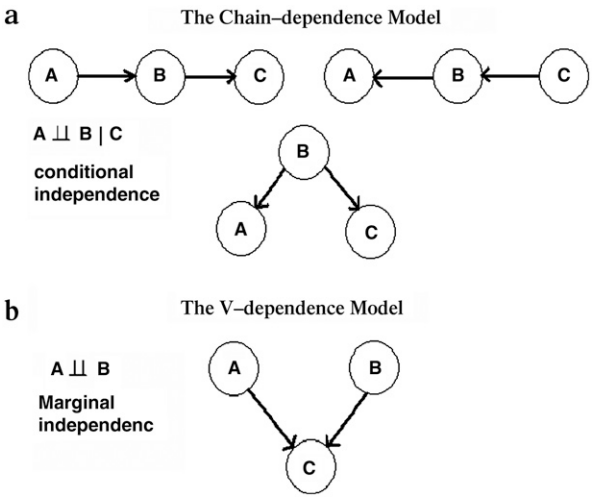

Fig. S6. Illustration of the chain-dependence model structure. The variables in set A are independent of the variables in set C conditional on the variables in set B. There are three different formulizing directions shown in the figure, but they are all equivalent (for a detailed proof, see Heckerman et al. (1995) (1). If set C is empty, the chain-dependence model is equivalent to the saturated model 1) (note that saturated model is also used for group 2 in BVP).

Illustration of the V-dependence model structure. The variables in set A are marginally independent of the variables in set C, but the variables in set B are dependent of the variables in both A and C. Thus, the variables in set A are NOT conditionally independent of the variables in set C.

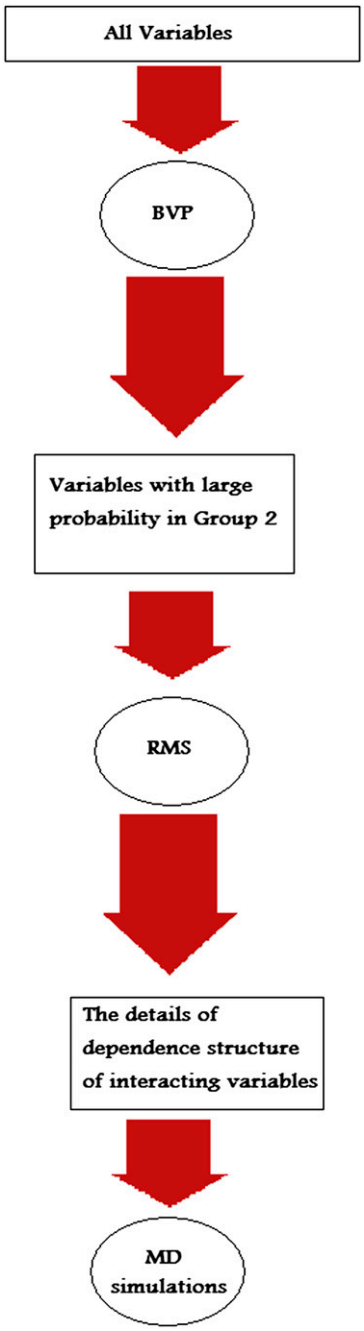

Fig. S7. The flow chart of the analysis pipeline we used in this paper.

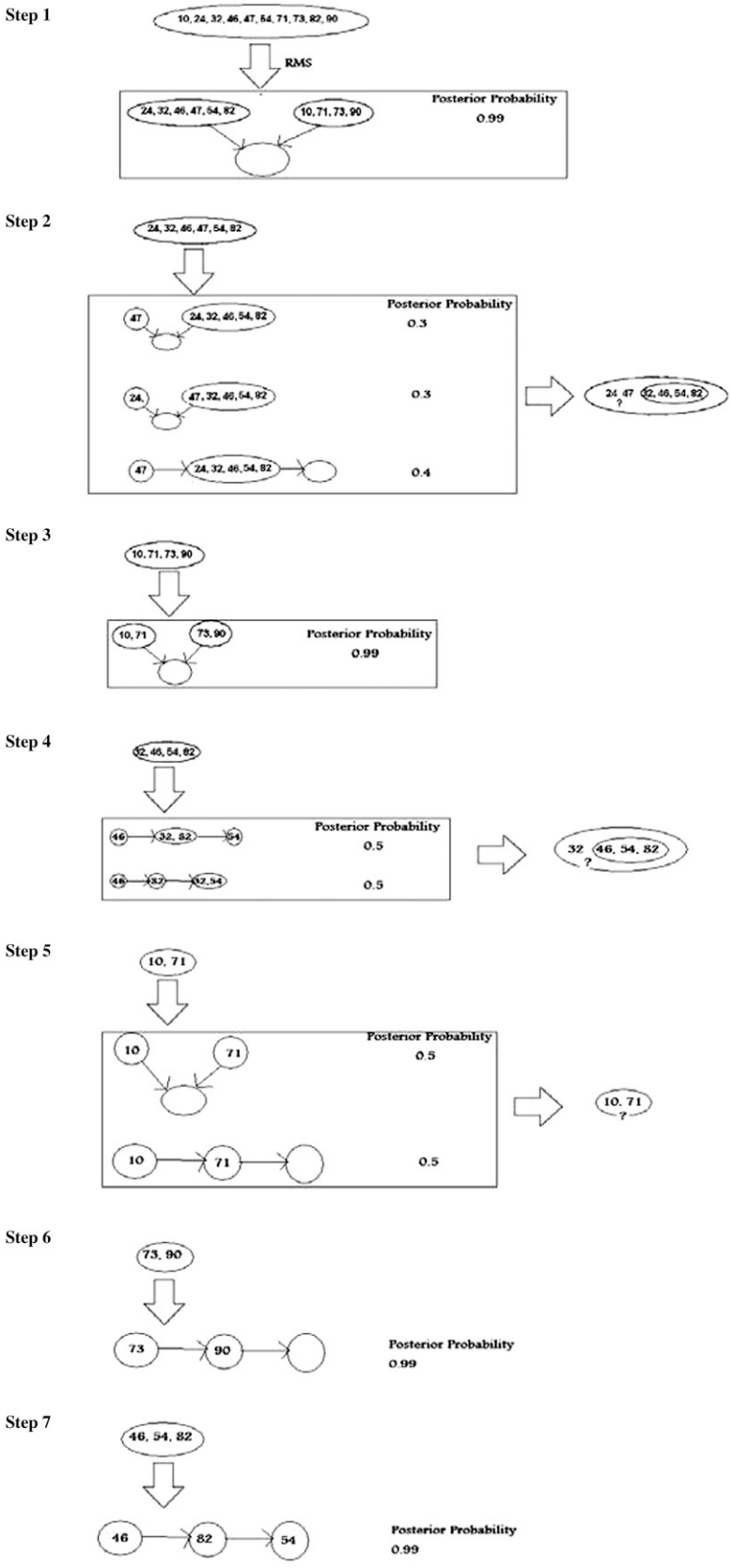

Fig. S8. The detailed steps of applying RMS to HIV drug resistant mutations 10, 24, 32, 46, 47, 54, 71, 73, 82, and 90.

Table S1. Mutation patterns that are found to be associated with indinavir treatment

Interaction patterns						Posterior prob.
24	32	46	54	82		0.571011
24	32	47	54	71	82	0.160826
32	46	54	82			0.126203
10	54	73	82	90		0.011326
46	54	82				0.010719
71	73	90				0.003396
10	54	82				0.001243
32	47	71	73	90		0.001
24	32	54	73	82	90	0.001
24	32	46	47	54	82	0.000833
73	90					0.000752
54	71	73	82	90		0.000675
73	90	95				0.000531
32	46	54	82	88		0.000397
32	47	54	71	82		0.000342
24	32	47	54	82		0.000145
24	32	54	71	82		0.000101

Interactions found to be associated with indinavir treatment. Shown here are the first 17 of them, with their respective posterior probabilities of association. The top interaction pattern has a posterior probability >0.5.

Table S2. Configurations of interaction pattern 24, 32, 46, 54, and 82 for indinavir treatment

24	32	46	54	82	Case (949 treated with INDINAVIR)	Control (4146 untreated)	chi square p.value (degree freedom=1)	N (Number of matching sequences in geno-pheno dataset)	Median (fold relative to wild type)	Q1 (1st quartile)	Q3 (3rd quartile)
I	V	L	V	A	0.018967	0	0	42	34.9	15.8	73.15
L	V	L	V	A	0.030558	0.000241	0	67	20.2	11.7	52.1
L	V	M	V	A	0.048472	0.000241	0	134	16.15	7.7	39
L	I	I	I	A	0.017914	0	0	18	16	8.05	28.65
L	V	I	I	V	0.070601	0.001206	0	181	12	4	39.7
L	V	L	I	A	0.020021	0	0	16	7.5	2.1	22
L	V	M	I	A	0.045311	0	0	33	4.7	1.8	10
L	V	M	I	V	0.537408	0.966956	0	1068	1	0.7	2
L	V	L	I	V	0.021075	0.001206	6.95E−14	35	4	2	10.25
L	I	L	I	A	0.014752	0	1.19E−13	8	9.8	6	21
L	V	I	V	A	0.013699	0.000241	2.27E−11	55	33.2	13	70.9
L	V	M	I	T	0.011591	0	9.83E−11	9	1	1	2
L	V	I	I	A	0.010537	0	9.23E−10	9	13	9.6	25
I	V	M	V	A	0.009484	0	8.68E−09	20	15.05	9	42.35
I	V	L	I	A	0.009484	0	8.68E−09	4	11.25	7.7	14.8
L	I	I	I	V	0.009484	0	8.68E−09	21	8	4.4	25.5
I	V	I	I	V	0.00843	0	8.23E−08	3	3.2	2.45	13.85
I	V	I	V	A	0.007376	0	7.83E−07	31	18	6.7	37.3
I	V	I	I	A	0.007376	0	7.83E−07	3	10	9	25.5
L	V	I	I	T	0.007376	0	7.83E−07	30	8.5	4	25
L	V	I	V	F	0.006322	0	7.53E−06	11	79	42.05	104.9
L	V	L	I	T	0.006322	0	7.53E−06	2	44.6	44.6	60
L	V	M	V	F	0.005269	0	7.30E−05	11	32	30	72.95
L	V	M	I	I	0.001054	0.024843	7.35E−05	33	0.8	0.4	1.3

Detailed configurations of the first interaction 24, 32, 46, 54, and 82. The "case" column is the frequencies of this configuration in 949 indinavir-treated sample; the "control" column is the frequencies of this configuration in 4146 untreated sample. The "chi square p.value" column shows the p-value for the χ^2 test (testing whether this configuration has different frequencies in cases and controls, d.f. $= 1$) after the Bonferroni correction. For those configurations that occurred fewer than 5 times in both cases and controls, we did not do the χ^2 test, thus, they are not shown in this table. "N" is the number of matching sequences in Stanford genotype-phenotype database. "Median" is the median of the fold resistance relative to the wild type, "Q1" and "Q3" are the 1st and 3rd quartiles, respectively. The configuration highlighted as red is the wild type configuration at these five positions 24, 32, 46, 54, and 82.

Table S3. Mutation patterns that are found to be associated with zidovudine treatment

Interactions						Posterior Prob
41	67	210	215	219		0.278588
41	67	70	210	215	219	0.267
41	67	69	70	210	219	0.262
41	67	210	219			0.063238
41	210	215				0.05279
41	210	215	219			0.048
41	67	70	210	219		0.028
41	44	210	215			0.00021
41	44	67	210	219		0.00011

Interactions found to be associated with zidovudine treatment. Shown here are the interactions with their posterior probabilities of association >0.0001.

Table S4. Configurations of interaction pattern 41, 67, 210, 215, 219 for zidovudine treatment

41	67	210	215	219	Case (339 treated with zidovudine)	Control (2187 untreated)	chi square p.value (degree freedom=1)	N (Number of matching sequences in geno-pheno dataset)	Median (fold relative to wild type)	Q1 (1st quartile)	Q3 (3rd quartile)
M	D	L	T	K	0.469027	0.951532	0	769	1	0.7	1.5
L	D	L	Y	K	0.088496	0.000457	0	136	5.4	2.3	18.9
L	D	W	Y	K	0.061947	0.000457	0	148	19.8	4.9	122
M	D	L	Y	K	0.061947	0.000457	0	60	4.9	1.5	22
M	N	L	T	Q	0.058997	0.000457	0	54	3.95	1.9	13.4
L	N	W	Y	K	0.029499	0.000457	6.23E−13	172	48.5	12	146
M	N	L	Y	Q	0.017699	0	7.04E−09	22	23	5	150
M	N	L	T	K	0.014749	0	1.96E−07	14	2.5	1.5	7.1
M	D	L	X	K	0.014749	0	1.96E−07	2	2	2	2
M	N	L	F	Q	0.020649	0.000914	2.09E−07	63	9.9	3.05	82.7
M	D	L	S	K	0.00885	0.002286	0.680554	1	3	3	3
M	D	F	T	K	0	0.005944	1	1	0.8	0.8	0.8
M	A	L	T	K	0	0.00503	1	1	2.1	2.1	2.1
L	D	L	T	K	0.0059	0.003201	1	15	2	0.65	4.5

Detailed configurations of the first interaction 41, 67, 210, 215, and 219 for zidovudine treatment. The "case" column is the frequencies of this configuration in 339 zidovudine-treated sample; the "control" column is the frequencies of this configuration in 2,187 untreated sample. "chi square p.value" shows the p-value for χ^2 test (testing whether this configuration has different frequencies in case and control, degree freedom equals 1) after Bonferroni correction. For those configurations that occurred fewer than five times in both case and control groups, we did not do the χ^2 test, thus they are not shown in this table. "N" is the number of matching sequences in the Stanford genotype-phenotype database. "Median" is the median of the fold resistance relative to the wild type, "Q1" and "Q3" are the 1st and 3rd quartiles, respectively. The configuration highlighted as red is the wild type configuration at these five positions 41, 67, 210, 215, 219.