nature genetics

A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase

Trevor Hinkley[1], João Martins[1], Colombe Chappey[2,3], Mojgan Haddad[2], Eric Stawiski[2,3], Jeannette M Whitcomb[2], Christos J Petropoulos[2] & Sebastian Bonhoeffer[1]

The development of a quantitative understanding of viral evolution and the fitness landscape in HIV-1 drug resistance is a formidable challenge given the large number of available drugs and drug resistance mutations. We analyzed a dataset measuring the *in vitro* fitness of 70,081 virus samples isolated from HIV-1 subtype B infected individuals undergoing routine drug resistance testing. We assayed virus samples for *in vitro* replicative capacity in the absence of drugs as well as in the presence of 15 individual drugs. We employed a generalized kernel ridge regression to estimate main fitness effects and epistatic interactions of 1,859 single amino acid variants found within the HIV-1 protease and reverse transcriptase sequences. Models including epistatic interactions predict an average of 54.8% of the variance in replicative capacity across the 16 different environments and substantially outperform models based on main fitness effects only. We find that the fitness landscape of HIV-1 protease and reverse transcriptase is characterized by strong epistasis.

With more than 20 drugs currently licensed to treat HIV infection[1] and over 200 mutations associated with drug resistance[2–5], it is increasingly difficult to develop a comprehensive understanding of HIV drug resistance. Resistance mutations differ in their potency to resist drug pressure[6,7], vary in their degree of cross resistance to different drugs or drug classes[8] and differ in the fitness costs induced in the absence of treatment[9–11]. Moreover, their effects depend to varying degrees on the context of accompanying mutations[7,12]. The quantitative dissection of the fitness effects of resistance mutations in the presence or absence of drugs and, in particular, the determination how the effect of mutations depends on the presence or absence of other mutations thus represents a major challenge.

The delineation of epistatic interactions between mutations is not only a matter of the size of the dataset. The combinatorial complexity of the genetic context in which any mutation appears explodes to a degree such that the estimation of the fitness effects is not feasible with standard statistical approaches, as the number of parameters to be estimated easily outnumbers the number of data points available even for the largest datasets. Problems in which the combinatorial complexity overwhelms standard methods of parameter inference are a common challenge in systems biology, and various approaches have been developed that allow reliable parameter estimation under conditions that lead to overfitting with standard statistical approaches. To overcome the problem of the large number of parameters and to account for non-normality in the error structure, we employed here generalized kernel ridge regression (GKRR), a regression method, which, in essence, penalizes against parameters that have low explanatory power. We used GKRR to quantify the fitness effects of amino acid variants using a dataset that measured *in vitro* fitness of 70,081 HIV-1 samples in the absence of drugs and in the presence of 15 different individual drugs. The samples were obtained from HIV-1 subtype B infected individuals undergoing routine drug-resistance testing (Online Methods). Our approach allows the reconstruction of an approximate fitness landscape of HIV protease and reverse transcriptase and thus offers the first quantitative description of a large, realistic and biologically relevant fitness landscape.

In vitro fitness of viral isolates is measured by replicative capacity. Viral isolates are sequenced in amino acids 1 to 99 of protease and 1 to 305 of reverse transcriptase (Online Methods). We quantified the fitness effects that are attributable to individual amino acid variants (main effects) and to pairwise epistatic effects between such variants (interactions) using GKRR. In particular we fitted two alternative models: (i) the ME model, which predicts fitness only on the basis of the main effects, and (ii) the MEEP model, which predicts fitness using both main effects and interactions. We applied GKRR because the size of the dataset used was too great for implementations of other regularization techniques such as the LASSO[13] or Dantzig selector[14].

Figure 1 shows the predictive power of the ME and MEEP models based on a sixfold cross validation by randomly subdividing the dataset into training and test sets of 65,000 and 5,000 independent virus samples, respectively. The goodness of the fit of the model is quantified by the percentage deviance explained (**Supplementary Note**, section 1.2). Deviance is the standard measure of goodness of fit in generalized models (that is, in models with non-normal error structure) and is analogous to the coefficient of determination R^2 of linear models with normal error structure[15]. The predictive power across the environments ranges for all cross validations from 35.0% to 65.9% for the MEEP model and from 26.8% to 57.9% for the ME model. The MEEP model has an average predictive power of 54.8%

[1]ETH Zürich, Institute of Integrative Biology, Zürich, Switzerland. [2]Monogram Biosciences, South San Francisco, California, USA. [3]Present address: Genentech, South San Francisco, California, USA. Correspondence should be addressed to S.B. (sebastian.bonhoeffer@env.ethz.ch) or C.J.P. (cpetropoulos@monogrambio.com).

Received 16 July 2010; accepted 3 March 2011; published online 27 March 2011; doi:10.1038/ng.795

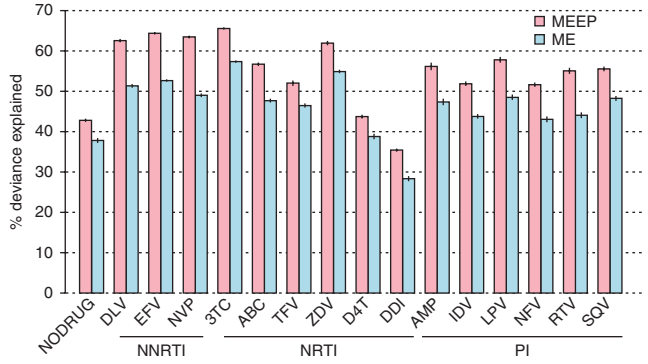

Figure 1 Analysis of predictive power. The figure shows the predictive power of the ME and MEEP models in a drug-free and 15 drug-containing environments. The predictive power is measured by the percentage deviance explained in a cross-validation dataset based on 5,000 independent virus samples. The bars represent mean, and the whiskers represent the standard errors from a sixfold cross validation. The MEEP model outperforms the ME model in all environments. The drugs used here are the protease inhibitors amprenavir (AMP), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), ritonavir (RTV) and saquinavir (SQV), the six nucleoside reverse transcriptase inhibitors abacavir (ABC), didanosine (ddI), lamivudine (3TC), stavudine (d4T), zidovudine (ZDV) and tenofovir (TFV), and the non-nucleoside reverse transcriptase inhibitors delavirdine (DLV), efavirenz (EFV) and nevirapine (NVP).

across all 16 environments. The MEEP model represents, on average, an 18.3% improvement in predictive power relative to the ME model. Note that in a regularized regression such as GKRR, increase in predictive power measured by cross validation is the appropriate model validation method. Hence, the substantial increase in predictive power of the MEEP over the ME model validates the inclusion of epistatic terms irrespective of their large number. Our kernelized approach allows including higher order epistatic interactions without substantial increases in computational requirements. Including three-way epistasis marginally decreases predictive power (data not shown). This decrease is due to the substantial increase in effective number of coefficients and does not imply that higher order epistatic interactions do not contribute to fitness.

We took an analogous approach to investigate the relative role of intragenic versus intergenic epistasis (interactions within protease or reverse transcriptase versus interactions between protease and reverse transcriptase). We fitted four models: main effects only (ME model), main effects + intragenic epistasis, main effects + intergenic epistasis and the full MEEP model (**Fig. 2**). Including intragenic epistasis consistently led to a much greater gain of predictive power than including intergenic epistasis. The main effects + intragenic epistasis model is generally as good, and sometimes even marginally better, than the MEEP model, which indicates that adding intergenic epistatic effects to the main effects + intragenic epistasis model does not further improve the predictive power. Decreases in predictive power are attributable to the fact that adding a large number of unnecessary parameters to a model can result in a reduction in predictive power in GKRR.

To verify that the estimates of the MEEP model are meaningful, we obtained sequences of protease and reverse transcriptase of treated and untreated patients from the Stanford HIV Drug Resistance Database[16] (see URLs) and determined the change of frequency of amino acid variants in treated versus untreated patients. The change of frequency of amino acid variants was significantly correlated with the fitness gain of amino acid variants in the presence compared to the absence of drugs relative to the consensus sequence ($P < 10^{-16}$ and Spearman rank correlation $\rho = 0.33$; Online Methods).

Because protein structure and epistasis are interrelated[17,18], we investigated the relation between epistasis in the drug-free environment and protease structure as an independent verification that the estimates of the 802,611 epistatic effects are biologically meaningful. **Figure 3** shows the strength of the epistatic effects between amino acid residues of the HIV-1 protease, revealing significant enrichment in epistatic interactions in the flap elbow, the cantilever and the fulcrum, structural units that have previously been described as being important to protein function[19]. Bootstrap analysis by random shuffling of the protein sequence revealed that epistasis is significantly enriched both within these structural domains and between the structural domains and the rest of the protein ($P < 10^{-5}$ for both tests; **Supplementary Fig. 1** and Online Methods). Moreover, in accordance with expectation, the strength of the epistatic interactions between amino acid residues correlates with physical proximity in the three-dimensional structure of protease ($P = 0.00857$, based on 100,000 bootstrap repeats; **Supplementary Fig. 2** and Online Methods). The significant correlation between epistasis and secondary structure or proximity shows that the estimated epistatic effects are biologically meaningful. Such correlations could not have been produced artifactually, as our procedure includes no structural information for parameter estimation.

Previous studies on epistasis in viruses did not allow a comprehensive quantification of individual fitness effects and epistatic interactions because they focused either on a limited set of interactions[20], made use of sequence data only[21] or did not correct for the effect of the genetic background[12]. Our study shows that despite the combinatorial complexity of the problem, biologically meaningful estimates for main effects and epistatic interactions can be obtained from large datasets that link fitness measurements to genotype. We verified the estimated effects using independent data. First, we showed that models including epistatic interactions explain on average 54.8% of the deviance in fitness across the 16 different environments based on sixfold cross validation. Second, we found a highly significant correlation between the change of the estimated main effects in the presence compared to the absence of drugs and the change in frequency of the corresponding amino acid variants in treated versus untreated patients based on independent data from the Stanford HIV Drug Resistance Database[16]. Finally, we found a correlation between epistasis and protease structural domains or physical proximity in the three-dimensional structure of protease.

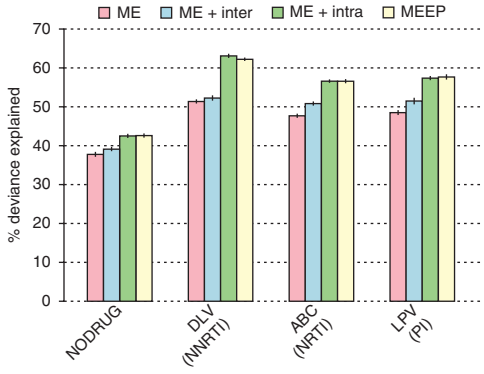

Figure 2 Analysis of predictive power of different epistatic models for four representative environments. The figure shows that most of the predictive power attributable to epistasis is in fact attributable to intragenic rather than intergenic epistatic interactions. In the non-nucleoside reverse transcriptase inhibitor environment, adding intergenic epistasis decreases predictive power. This decrease reflects that adding a large number of parameters with little or no explanatory power can reduce the predictive power of GKRR. The bars represent mean and the whiskers the standard errors from a sixfold cross validation.

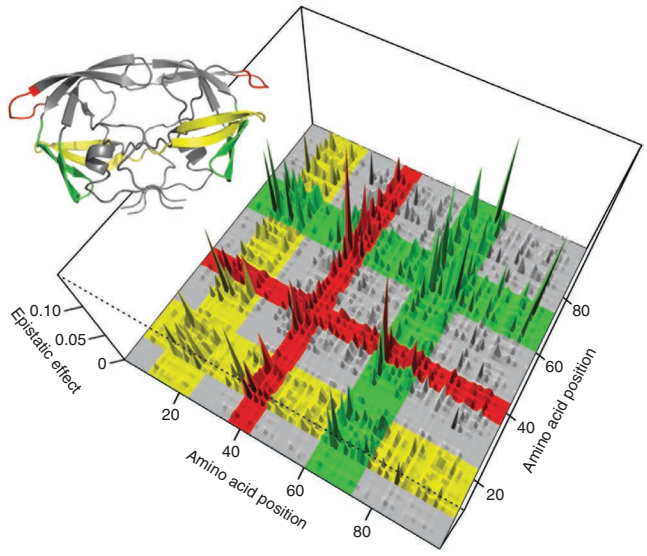

Figure 3 Cumulative strength (CS) of the absolute epistatic effects in the HIV-1 protease as measured in the drug-free environment. The cumulative effect between two positions is calculated as the sum over the absolute values of all epistatic interactions between the amino acid variants at those positions as estimated by the MEEP model. We plotted $CS^{1.5}$ to enhance visual clarity. The regions corresponding to the flap elbow, fulcrum and cantilever, colored in red, yellow and green, respectively, are significantly enriched in epistasis (**Supplementary Fig. 1**). The inset shows the structure of the HIV-1 protease (Protein Data Bank ID 1A30, rendered with PyMOL; see URLs). The region enriched in epistatic interaction, corresponding to the flap elbow, is somewhat larger than the literature description of this region[19].

Ever since the synthesis of Darwinian evolution with genetic inheritance in the early 20th century, the debate about the relative role of epistasis and main effects in determining fitness has remained at the heart of evolutionary genetics[22,23]. With the advent of systems biology, it has become possible to measure these epistatic effects more comprehensively[24–28]. Supporting Sewall Wright's view of the dominant role of epistasis[22,23], we find that epistasis and, in particular, intragenic epistasis, is crucial in determining fitness. For our dataset, the inclusion of epistatic interactions improved the predictive power by an average of 18.3% across all environments. Our approach provides us with a predictive model for realistic fitness landscapes, opening up new avenues to study evolutionary adaptation on complex fitness landscapes and to simulate the evolution of drug resistance.

URLs. Stanford Drug Resistance Database, http://hivdb.stanford.edu; Pymol, http://www.pymol.org/; HIPAA practices of Monogram Biosciences, http://www.monogrambio.com/990HIPAA.aspx; the data management plan is found at http://precedings.nature.com/documents/5668/version/1; data requests can be sent through http://www.monogrambio.com/860ResearchAndDevelopment.aspx.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

T.H., J.M. and S.B. thank the Swiss National Science Foundation (NF 3100A0-116408) for financial support. We thank R. Regös, R. Kouyos, J. Engelstädter, S. Alizon and T. Gernhard-Stadler for valuable comments and critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

T.H. developed and implemented the generalized kernel ridge regression and analyzed data. J.M. analyzed data. C.C., M.H., E.S., J.M.W. and C.J.P. generated and pre-processed the experimental data. S.B. designed the study and analyzed data. T.H. and S.B. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. De Clercq, E. Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *Int. J. Antimicrob. Agents* **33**, 307–320 (2009).
2. Shafer, R.W. & Schapiro, J.M. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev.* **10**, 67–84 (2008).
3. Clavel, F. & Hance, A.J. HIV drug resistance. *N. Engl. J. Med.* **350**, 1023–1035 (2004).
4. Johnson, V.A. *et al.* Update of the drug resistance mutations in HIV-1. *Top. HIV Med.* **16**, 138–145 (2008).
5. Bennett, D.E. *et al.* Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE* **4**, e4724 (2009).
6. Petropoulos, C.J. *et al.* A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **44**, 920–928 (2000).
7. Rhee, S.-Y., Liu, T., Ravela, J., Gonzales, M.J. & Shafer, R.W. Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrob. Agents Chemother.* **48**, 3122–3126 (2004).
8. Harrigan, P.R. & Larder, B.A. Extent of cross-resistance between agents used to treat human immunodeficiency virus type 1 infection in clinically derived isolates. *Antimicrob. Agents Chemother.* **46**, 909–912 (2002).
9. Croteau, G. *et al.* Impaired fitness of human immunodeficiency virus type 1 variants with high-level resistance to protease inhibitors. *J. Virol.* **71**, 1089–1096 (1997).
10. Martinez-Picado, J., Savara, A.V., Sutton, L. & D'Aquila, R.T. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *J. Virol.* **73**, 3744–3752 (1999).
11. Mammano, F., Trouplin, V., Zennou, V. & Clavel, F. Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug. *J. Virol.* **74**, 8524–8531 (2000).
12. Bonhoeffer, S., Chappey, C., Parkin, N.T., Whitcomb, J.M. & Petropoulos, C.J. Evidence for positive epistasis in HIV-1. *Science* **306**, 1547–1550 (2004).
13. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**, 407–499 (2002).
14. Candes, E. & Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* **35**, 2313–2351 (2007).
15. Nelder, J. & Wederburn, R. Generalized linear models. *J. Roy. Stat. Soc. A* **135**, 370–384 (1972).
16. Shafer, R.W. Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* **194** Suppl 1, S51–S58 (2006).
17. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
18. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
19. Hornak, V., Okur, A., Rizzo, R. & Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **103**, 915–920 (2006).
20. Sanjuán, R., Moya, A. & Elena, S.F. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc. Natl. Acad. Sci. USA* **101**, 15376–15379 (2004).
21. Chen, L., Perlina, A. & Lee, C.J. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.* **78**, 3722–3732 (2004).
22. Provine, W. *The Origins of Theoretical Population Genetics* (The University of Chicago Press, Chicago, Illinois, USA, 1971).
23. Wolf, J., Brodie, E. III & Wade, M. *Epistasis and the Evolutionary Process* (The University of Chicago Press, Chicago, Illinois, USA, 2000).
24. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
25. Yeh, P.J., Hegreness, M.J., Aiden, A.P. & Kishony, R. Drug interactions and the evolution of antibiotic resistance. *Nat. Rev. Microbiol.* **7**, 460–466 (2009).
26. Jasnos, L. & Korona, R. Epistatic buffering of fitness loss in yeast double deletion strains. *Nat. Genet.* **39**, 550–554 (2007).
27. Kouyos, R.D., Silander, O.K. & Bonhoeffer, S. Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol. Evol.* **22**, 308–315 (2007).
28. de Visser, J.A.G.M. & Elena, S.F. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.* **8**, 139–149 (2007).

ONLINE METHODS

Data. We obtained 70,081 virus samples from HIV-1 subtype B infected individuals undergoing routine drug resistance testing[29]. These data were collected and used in agreement with HIPAA practices (for further information, see HIPAA practices of Monogram Biosciences and URLs). The samples were assayed for replicative capacity based on the construction of HIV-derived test vectors. The assay to measure viral replicative capacity in absence of drugs has been described in detail elsewhere[6]. In brief, patient virus derived amplicons representing all of protease and most of reverse transcriptase are inserted into the backbone of a resistance test vector. This vector is based on the NL4-3 molecular HIV clone and has been modified such that it can only undergo a single round of replication. The replicative capacity assay then quantifies the total production of infectious progeny virus after a single round of infection of the patient-derived virus relative to that of an NL4-3 based control virus. The replicative capacity of the NL4-3–based control virus thus equaled 1. The replicative capacity measures the total reproductive output relative to a control virus in a single round of replication and can thus be regarded as a proxy for viral fitness[30].

Commonly, the replicative capacity is measured in the absence of drugs. For the virus samples analyzed here, the replicative capacity was also measured in the presence of 15 different single drugs at a series of drug dilutions. The drugs used here were the protease inhibitors amprenavir (AMP), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), ritonavir (RTV) and saquinavir (SQV), the six nucleoside reverse transcriptase inhibitors abacavir (ABC), didanosine (ddI), lamivudine (3TC), stavudine (d4T), zidovudine (ZDV) and tenofovir (TFV), and the non-nucleoside reverse transcriptase inhibitors delavirdine (DLV), efavirenz (EFV) and nevirapine (NVP). For each drug, the replicative capacity of a virus on drugs was given by the interpolated value measured at the drug concentration at which the NL4-3–based control virus has 10% of its replicative capacity in the absence of drug (the IC90 for NL4-3 was used as the reference drug concentration for every subsequent measurement). In addition to the fitness measurement on and off drugs, all of the protease and the amino acids 1 to 305 of reverse transcriptase were sequenced by population sequencing for all virus samples included in this analysis.

Note that replicative capacity is different from IC50 and EC50, other commonly used phenotypic measures of drug resistance which measure the drug concentration at which a virus sample is half maximally inhibited. Previous algorithms to predict phenotypic properties of drug resistance have focused on the prediction of IC50 (ref. 31). By measuring a drug concentration that causes a relative change in activity, IC50 discards information about the absolute fitness. Replicative capacity, however, does not measure a change in activity but an absolute activity at a given drug concentration (previously measured as the IC90 of the reference NL4-3). Replicative capacity, therefore, is a more appropriate measure of viral fitness. However, because replicative capacity measures absolute activity, it is a more complex phenotypic measure and therefore harder to predict. To test this statement, we also tested our algorithm against a measure similar to IC50, defined by replicative capacity in presence of drugs relative to the corresponding replicative capacity in absence of drugs. This simpler fitness resulted in an average predictive power of 89% and a maximum predictive power of 95% across all the drug environments.

Amino acid sequences of the protease gene and the partial reverse transcriptase gene were obtained by population sequencing for all virus samples included in this analysis[6]. Because of this population sequencing, sequences are defined in terms of probabilities of allele occurrences for each locus. To ease computational issues, we did not include any variant that appeared fewer than ten times in the entire dataset. The effect of this thresholding on predictive power was less than 0.01%. In our sequence data, there are N_M = 1,859 amino acid variants above the threshold and N_E = 802,611 pairwise combinations of these variants (which is a small subset of the theoretically possible set of combinations). If main effects or interactions always occur with other main effects or interactions, the effect that is attributable to the linked group is distributed evenly over all these coefficients as a result of the ridge regression methodology employed. Analysis of our data shows that we have altogether 659,654 independent effects.

Model fitting. We quantified the fitness effects that are attributable to individual amino acid variants independent of the genetic context (main effects) and the fitness effects attributable to pairwise epistasis between variants (interactions) by fitting the following model:

$$\log(W_i) = I + \sum_{j=1}^{N_M} M_{ij}\gamma_j + \sum_{k=1}^{N_E} E_{ik}\chi_k$$

Here, W_i is the replicative capacity (fitness) of sequence i. I is the intercept, which represents the log fitness of the NL4-3 reference sequence. The parameter γ_j represents the main effect of the j^{th} variant and M_{ij} is the probability of that variant occurring in a randomly selected sequence from the population i. Similarly, χ_k represents the interaction of the k^{th} combination of variants and E_{ik} is a variable that accounts for the presence or absence of that combination of variants in the sequence. The ME model uses only the 1,859 M_{ij} terms to compute predicted fitness and the MEEP model adds 802,611 E_{ik} terms to this model. These models are explained in depth in the **Supplementary Note**, section 2.

The model is fitted by generalized kernel ridge regression (GKRR), a technique that combines the fitting of non-normal error structure by the generalized linear model (GLM) with the capability of kernel ridge regression to fit data with fewer observations than dimensions. We give an intuitive introduction to our methodology in the **Supplementary Note**, section 1.1. A detailed technical explanation is provided in the **Supplementary Note**, sections 1.2 through 1.5, and **Supplementary Figure 3**. The software is available on request.

Statistical analysis. The change of frequency of single amino acid variants in HIV-1 protease and reverse transcriptase was determined based on 44,119 sequences obtained from the Stanford HIV Drug Resistance Database[16] derived from treated and untreated patients (numbers of sequences: reverse transcriptase, treated = 7,232; reverse transcriptase, untreated = 12,022; protease, treated = 10,011; protease, untreated = 14,854; downloaded: 17/09/2009, see URLs). The fitness gain was estimated as the difference between the maximal beneficial fitness effect of an amino acid variant in presence of drugs versus the fitness effect in absence of drugs. Note that fitness effects in different environments are correlated with drug class being a dominant factor[32]. Fitness effects of the amino acid variant were measured relative to the consensus amino acid variant in untreated patients. The significance of the correlation between fitness gain in presence versus absence of drugs and frequency change in treated versus untreated patients was calculated based on a Spearman rank correlation (N = 1,169 amino acid variants, $P \leq 10^{-16}$ and Spearman's ρ = 0.33).

To test for statistical significance of correlations between epistatic effects and protein structure, we used bootstrapping. To this end, we generated bootstrapped matrices of epistatic interactions by shuffling rows and columns of the estimated epistatic interaction matrix. We used 100,000 bootstraps to test to infer statistical significance of the enrichment of epistatic interactions within HIV-1 protease structural domains and between these structural domains and the remainder of the protein. We used 100,000 bootstraps to test to infer statistical significance of the Spearman rank correlation coefficient between strength of epistatic interactions between amino acid residues and their physical proximity in the 3D structure of protease.

To obtain values and standard errors for predictive power, sixfold cross validation was performed. For each cross-validation we selected two subsets of data from our database. The larger dataset, consisting of 65,000 sequences and corresponding fitness values, was used to estimate main effects and interactions. The smaller dataset, consisting of 5,000 sequences and fitness values, was not used for model fitting but was reserved only for the purpose of quantifying the goodness of the model fit in terms of the percentage deviance explained.

Data access. The data underlying this study can be accessed by submitting requests to Monogram Biosciences (see data requests, URLs). Access to these data is restricted to *bona fide* researchers under conditions specified in the data management plan[29] (URLs).

29. Petropoulos, C.J. Data management plan for Monogram BioSciences. *Nature Precedings* <http://dx.doi.org/10.1038/npre.2011.5668.1> (2011).
30. Dykes, C. & Demeter, L.M. Clinical significance of human immunodeficiency virus type 1 replication fitness. *Clin. Microbiol. Rev.* **20**, 550–578 (2007).
31. Rhee, S.-Y. *et al.* Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA* **103**, 17355–17360 (2006).
32. Martins, J.Z. *et al.* Principal component analysis of general patterns of HIV-1 replicative fitness in different drug environments. *Epidemics* **2**, 85–91 (2010).

doi:10.1038/ng.795