Figures and figure supplements

Epistasis and entrenchment of drug resistance in HIV-1 subtype B

Avik Biswas et al
Figure 1. The Potts model predicts residue frequencies. (A) Schematic showing that the Potts model can be used to classify sequences by how likely a residue $a$ is to appear at a position $i$ in a sequence $S$ using the background-dependent probability, $P(S, i, a)$. (B) The observed frequency of the resistance mutation L90M in HIV-1 drug-experienced proteases matches the Potts model predicted frequencies in sequence clusters binned according to the Potts-predicted frequencies in steps of 0.1 (blue circles) with statistical error (green). Diameters of the circles represent the number of sequences. Inset shows the significant overlap in Hamming distance for sequences with low predicted mutant frequencies between 0.2 and 0.3 (blue) and high, 0.7 and 0.8 (pink) depicting the difficulty of such a classification based on Hamming distance. (C) The receiver operator characteristic (ROC) curve comparing the Potts model and Hamming distances as classifiers of mutational probabilities for L90M in HIV protease. (D) The average absolute error between the observed mutational frequencies and the Potts-predicted frequencies for the major drug-resistance mutations in HIV-1 in three drug targets. The average absolute error is calculated by binning the sequences in ascending order of their predicted frequencies such that there are roughly equal number of sequences in each bin as shown in Figure 1—figure supplement 1, and averaging over the absolute error in each bin.

DOI: https://doi.org/10.7554/eLife.50524.002
Figure 1—figure supplement 1. Observed vs predicted frequencies for calculating the average absolute error in Figure 1D. The average absolute error in Figure 1D is calculated by binning the sequences in ascending order of their predicted frequencies such that there are roughly equal number of sequences in each bin, and averaging the absolute error of the bins. This binning procedure avoids finite sampling errors encountered when following a more restrictive method of binning by Potts frequencies from 0 to 1 in steps of 0.1 as shown in Figure 1B especially when the mutation frequency is small. The observed vs predicted frequencies are shown for primary drug-resistance mutations V82A occurring in response to PIs in HIV-1 protease (A); N155H occurring in response to INSTIs in HIV-1 IN (B); M184V occurring in response to NRTIs in HIV-1 reverse transcriptase (C); and Y181C/G occurring in response to NNRTIs in HIV-1 reverse transcriptase (D). Diameters of the blue circles represent data sizes in each bin, and error bars are shown as green cross-hairs.

DOI: https://doi.org/10.7554/eLife.50524.003
Figure 2. Entrenchment and the effect of epistasis on the favorability of a primary resistance mutation. (A) $\Delta E^{90} = E_{\text{wild-type}}^{90} - E_{\text{mutation}}^{90}$, change for sequences, conditional on the number of PI-associated mutations is shown as boxplots annotating the first, second and third interquartile range. Whiskers extend to 1.5 times the interquartile range with outliers marked as ‘x’s and the mean values are marked as squares. The left ordinate scale shows the relative probability of reversion ($e^{-\Delta E}$), and the right shows $\Delta E$. Sequences whose energy difference fall above $\Delta E = 0$ (dashed line) are entrenching backgrounds favoring the mutation. Sequence backgrounds where the mutation is favored on average are shown in red, the others in blue. The mutation L90M becomes favorable on average when there are about 9 PI-associated mutations, but there is a wide range of favorability and ‘which’ PI-associated mutations are present play an important role in determining if the primary mutation is favored/disfavored. The highlighted regions (white with dark border) show there are many sequence backgrounds with between 7 and 14 mutations in which L90M is either ‘highly entrenched or favored’ (top) or ‘highly disfavored’ (bottom). (B) Distributions of number of PI-associated mutations in sequences in the ‘highly entrenching’ and ‘highly disfavoring’ regions from panel A have a large overlap, again showing that entrenchment is not primarily determined by number of associated mutations. Prediction of the likelihood of M based on the number of PI-associated mutations alone for these sequence backgrounds where M is highly entrenched (red) or highly disfavored (blue) would be especially difficult due to the large overlap between them.

DOI: https://doi.org/10.7554/eLife.50524.004
Figure 3. Degree of entrenchment of key resistance mutations occurring in the catalytic core domain (CCD) of HIV-1 IN. The change in Potts statistical energy $\Delta E$ for some of the key resistance mutations occurring in the catalytic core domain (CCD) of IN, is plotted as a function of the rank of mutation-carrying sequences, ranked in descending order of their favorability towards the mutant. Plot shows the degree of entrenchment for these mutations. For example, Q148R is highly entrenched in almost all sequences carrying it, whereas, G163K is entrenched ($\Delta E > 0$) in only about half of the sequences carrying G163K.

DOI: https://doi.org/10.7554/eLife.50524.011
Figure 4. Distribution of Potts $\Delta E$ scores for key residues associated with drug resistance in HIV-1 IN. The distribution of the Potts $\Delta E(E_{\text{wild}} - E_{\text{mut}})$ scores for sequences carrying the particular resistance mutation are shown in ‘green’ for the most frequently observed INSTI selected resistance mutations in HIV IN, and in ‘blue’ for all other possible mutations at the same sites. Other possible mutations include rarely observed or unobserved mutations. The histograms show the differential distribution of $\Delta E$ scores for observed vs. unobserved/rare mutations at 15 primary mutation sites associated with evolving drug resistance in HIV-1 IN. The green (observed) and blue (unobserved/rare) distributions are normalized to the total number of primary DRMs in IN in the Stanford HIVDB and the total number of other possible mutations at the same sites, respectively. The mean $\Delta E$ scores for observed vs. unobserved mutations are +2.11 and −5.58, respectively ($p<0.001$). The wide distribution of $\Delta E$ scores also illustrates the role of the background in which the resistance mutation occurs.

DOI: https://doi.org/10.7554/eLife.50524.013
Figure 5. Entrenchment and favorability of key resistance mutations in specific backgrounds. The $\Delta E$ change in Potts energy of a sequence is used as the measure of ‘entrenchment’ and favorability of key resistance mutations in HIV-1 NL4-3, HXB2, and the subtype-B consensus sequences, respectively, shown as a function of the average $\Delta E$ in drug-experienced HIV-1 subtype B patient populations ($\langle \Delta E_{\text{patient}} \rangle$) in the Stanford HIVDB for viral integrase (A,B,C) and reverse transcriptase (D,E,F). In each case, the Pearson correlation coefficients are indicated. The protein sequences for the molecular clones NL4-3 and HXB2 are obtained from GenBank with accession number AF324493.2 and K03455.1, respectively with protein ids for the pol polyprotein as AAK08484.2 and AAB50259.1, respectively. The subtype B consensus sequence is obtained from the Stanford HIVDB. The degree of ‘entrenchment’ in these subtype B strains is often not representative of the average entrenchment effects in a patient population or even the most representative background from a patient population. DOI: https://doi.org/10.7554/eLife.50524.014
Figure 6. The particular sequence background in which a resistance mutation occurs affects the degree of entrenchment, with often a clear distinction between sequences where the mutation is present (green) versus absent (blue) shown here for the mutations N155H (A) and G140S (B) in integrase. The degree of entrenchment for subtype B consensus, NL4-3 and HXB2 are shown as ‘black dashed’, ‘red’ and ‘magenta’ lines, respectively. The Potts entrenchment score manifests as a clear distinction between backgrounds where a particular mutation is observed from ones where the mutation is absent, with the former more likely to present a distinct fitness advantage towards the mutation.

DOI: https://doi.org/10.7554/eLife.50524.015
Figure 7. Comparison of Potts models parameterized on drug-naive and drug-experienced HIV protein sequences. The comparison of the effects of point mutations is shown in terms of Potts ΔE scores (which forms the basis of our study) using two different Potts models, parameterized on drug-naive vs drug-experienced sequences for PR drug-resistance mutations M46I (A), I54V (B), V82A (C), and L90M (D), all of which appear with at least a frequency of 0.25% in both datasets. Bin shading for the 2D histogram scatter plots shown here scales logarithmically with the number of sequences whose scores fall into each bin. To obtain ΔE scores, the sequences are scored using a drug-naive model vs. a drug-experienced model. The ΔE scores are highly correlated with a Pearson correlation coefficient of 0.82 (p<0.001), 0.93 (p<0.001), 0.85 (p<0.001), and 0.82 (p<0.001), respectively. DOI: https://doi.org/10.7554/eLife.50524.016
Figure 8. Drug-pressure associated mutations are largely common between drugs of the same class. The mutations (both primary and associated) arising in drug-treated HIV proteases in response to inhibitor treatment are shown corresponding to each protease drug. The diameters of the circles represent the number of mutations at that site that occurred in sequences treated with the particular drug. Most mutations that occur in response to one drug are seen to have occurred when treated with another drug of the same class, showing that the 'spurious correlations' (that could be picked up by a Potts model built on a mixture of patient sequences treated with different drugs of the same class, if the mutations occurring in response to one drug are not at all observed in response to another), are minimal.

DOI: https://doi.org/10.7554/eLife.50524.017
Residue positions are colored by respective domains and subdomains of the p66 subunit:

**Polymerase:**
- Fingers (1-85, 118-155)
- Palm (86-117, 156-236)
- Thumb (237-318)
- Connection (319-426)

**RNase H:**
- Fingers (1-85, 118-155)

Appendix 1—Figure 1. Sequence coverage for RT. Figure shows the sequence coverage (# of sequences vs the # of residues) for RT drug-experienced (both NRTI and NNRTI) sequences derived from the Stanford HIVDB (22,444 isolates from 20422 patients). For RT, sequences with exposure to both NRTIs and NNRTIs were selected as an alternate search for RT sequences exposed to only NRTIs or NNRTIs would return a vastly smaller number of isolates (5398 and 80, respectively). Sequences with insertions ('#') and deletions ('~') are removed. MSA columns and rows with more than 1% gaps ('.') are removed. This resulted in a final MSA size of $N = 19194$ sequences from 17130 persons each with length $L = 188$ for RT. To retain enough sequence coverage in the MSA, we removed residues: residues 1–38, and residue 227 onwards for RT. For this reason, some interesting DRMs like F227I/L/V/C, L234I, P236L or N348I (NNRTI affected) for RT are not amenable for our analysis.

DOI: https://doi.org/10.7554/eLife.50524.020