Measuring Selection Across HIV Gag: Combining Physico-Chemistry and Population Genetics

Elizabeth Johnson [1,2] and Michael A. Gilchrist* [2,3]

1 Microbiology, University of Tennessee, Knoxville, TN, United States

2 National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, United States

3 Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, United States

* Corresponding author: mikeg@utk.edu

Abstract

We present a quantitative, population genetics based physico-chemical model which predicts the stationary probability of observing an amino acid residue based on the optimal residue for the site and the sensitivity of the protein functionality to deviation from the optimum. We contextualize our physico-chemical model by comparing it to the more general, but less biologically meaningful models of sequence entropy. To illustrate our model's use, we parameterize our model using over a 1000 different sequences of HIV subtype C's Gag poly-protein. Using data from the LANL HIV database, we evaluate our physico-chemical model's performance by first comparing its site sensitivity parameters G' to the entropy based measures of site conservation and its ability to predict empirical *in vitro* and *in vivo* measures of HIV fitness. While our model's G' is well correlated with conservation, G' does a significantly better job predicting the empirical fitness data. More importantly, unlike the entropy model, our model can be further refined and used to test more complex biological hypotheses. For example, in our analysis we find evidence that different protein regions of the gag poly-protein have different sensitivities to deviation from the optimal amino acid residue's molecular

volume. Finally, given its biological basis, it should be possible to extend our method to include epistasis in a more realistic manner than Ising models while requiring many fewer parameters than Potts models.

Author Summary

TO BE WRITTEN

Introduction

HIV's protein sequence has been described as having a great deal of robustness or 'genetic plasticity' (Lemey et al., 2006; Salemi, 2013; Cuevas et al., 2015). This is because amino acid substitutions at many different sites appears to have little or no consistent effect on HIV fitness (Rihn et al., 2013a). HIV's genetic plasticity allows the viral populations within a host to evade or 'escape' a patient's immune response by simply evolving to a new location in epitope space with little or no fitness cost. HIV's ability to escape the host's immune response has impeded efforts to create an effective vaccine (Goulder and Watkins, 2004; Autran et al., 2008; Johnston and Fauci, 2008). In response, researchers have tried to identify amino acid sites under strong and consistent purifying selection (Ferguson et al., 2013; Barton et al., 2016). Sites experiencing strong which have the potential to be promising vaccine targets (Rolland et al., 2013).

In order to identify amino acid sites under consistent purifying selection, researchers have attempted to estimate HIV's fitness landscape (Deforche et al., 2008; Seifert et al., 2015; Kouyos et al., 2012; Hinkley et al., 2011; Shekhar et al., 2013; Ferguson et al., 2013; Mann et al., 2014; Lorenzo-Redondo et al., 2014; Moradigaravand et al., 2014; Barton et al., 2016). Since first introduced by Sewall Wright in 1932 (Wright, 1932), fitness landscapes have become a conceptual cornerstone within the field of evolutionary biology (e.g. Lande and Arnold, 1983; Lande, 1985, 1986; Kauffman and Levin, 1987; Charlesworth and Rouhani, 1988; Kauffman, 1993; Niklas, 1994; Gavrilets, 1997; Fontana, 2002; Berg and Lässig, 2003; Gavrilets, 2004; Berg et al., 2004; Wilke and Drummond, 2006; Gilchrist, 2007; Lässig, 2007; Calcott, 2008; Mustonen and Lässig, 2009; Draghi et al., 2010; Gilchrist et al., 2009; Wallace et al., 2013; Gilchrist et al.,

2015). In the simplest scenarios where the evolutionary process has either reached stationarity or we have no a priori information, the probability of observing a sequence i is proportional to its evolutionary fitness W_i raised to the effective population size N_e, i.e. $p_i \propto W_i^{N_e}$ (Wright, 1969; Iwasa, 1988; Berg and Lässig, 2003; Sella and Hirsh, 2005; McCandlish et al., 2015).

While often unaware of its theoretical foundation, HIV researchers have extensively used this link between the fitness contribution of an amino acid residue at a particular site and its observed frequency (e.g. Rihn et al., 2013a; Liu et al., 2006). In one such approach, researchers have adapted the Potts and Ising models from statistical mechanics to quantify direct and epistatic fitness effects between amino acid sites (Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2016). While the Ising model, which categorizes amino acid residues as simply either being optimal or non-optimal, has been criticized as being overly simplistic, it has been effectively employed to identify epistatic interactions between sites (Ferguson et al., 2013; Mann et al., 2014). In contrast, researchers have found the Potts model, which uses 20 categories per site and, in turn, require hundreds of epistatic terms between any two sites, difficult to fit to even large data sets (Barton et al., 2016).

Using another approach, researchers have used Shannon entropy E, specifically the inverse entropy of a site $1/E$ or its 'conservation', as a measure of the strength of purifying selection for the presumably optimal consensus amino acid residue (Dietrich and Skipper, 2012; Acevedo et al., 2014). In this approach, sites with little variation in amino acid usage have high $1/E$ values and are, in turn, inferred to be experiencing strong and consistent purifying selection. In contrast, sites with substantial variation have low $1/E$ values and are inferred to be either under weak or variable purifying selection. As we show later, the conservation of a site $1/E$ is proportional to the expected log-likelihood of observing a randomly chosen amino acid under a saturated, multinomial model parameterized with a given data set. Equivalently, $1/E$ can also be viewed as a summary statistic quantifying the ruggosity of HIV's fitness landscape (Kauffman, 1993).

Another shortcoming of using $1/E$ as a measure of consistent purifying selection is that it treats all non-optimal amino acid residues equally dissimilar from the optimal residue. This is undersirable given that the physico-chemical properties of amino acids

clearly affect the probability one amino acid will substitute for another (MIKE: citations from neutrality argument, David Liebels lab,phylogenetic studies Grantham, 1974; Wilke and Drummond, 2010). By treating amino acid residues as equally different, $1/E$ ignores the fact that amino acid residues have differing degrees of physico-chemical dissimilarity. As a result, $1/E$ has the potential to miss sites where there is strong, consistent purifying selection for a set of physico-chemical properties, but where these properties can be reasonably satisfied by more than one amino acid residue. More generally, the fact that $1/E$ ignores the physico-chemical properties of amino acid residues suggests it is glossing over potentially important information embedded within the data used to estimate $1/E$.

Despite the fact that site conservation $1/E$ ignores the physico-chemical characteristics and is not actually a measure of consistent purifying selection against non-optimal amino acid residues, it is, nevertheless, widely used. One contributing factor $1/E$ wide use is the fact that it can be easily calculated from sequence alignment data, c.f. the Potts model. Another contributing factor is the fact that $1/E$ has proved to be a useful predictor of sites with large impacts on HIV fitness.

In an effort to overcome the aforementioned shortcomings, we introduce a model that includes physico-chemical weighting terms and explicitly estimates the sensitivity of virus fitness G' to deviations in the physico-chemical properties of the optimal amino acid residue for that site. While the site sensitivity parameters G' we estimate correlate well with site conservation, G' does a better job predicting HIV genotype fitness from both *in vivo* and *in vitro* studies. Unlike site conservation $1/E$, our model allows us to test biologically motivated hypotheses. For example, we find that our physico-chemical weightings and the distributions of G' vary between the different protein regions of Gag.

Taken together, our physico-chemical model helps advance researchers ability to extract information from observational data and represents a biologically grounded framework which can be further extended and used to test clearly posed hypotheses. In its current form, our physico-chemical model only considers site independent effects and can be viewed as a more constrained or parsimonious version of the more general, parameter saturated, entropy model underlying the calculation of $1/E$. Given the parallels between evolution and statistical mechanics (Sella and Hirsh, 2005), it should be possible to extend our physico-chemical model to include epistatic effects in a more

realistic manner than the Ising models, but in a substantially more efficient manner, in terms of the number of parameters and the ease of parameter estimation, than the Potts models.

Methods

We begin by defining our physico-chemical model and, in the process, clearly define our site sensitivity parameter G'. Next, we review how site conservation $1/E$ is defined using the Shannon entropy model. As a result, we clearly show the link between the physico-chemical model and the entropy model and their corresponding probabilities of observing each amino acid residue at a particular site. Next, we describe the data and methods we used to parameterize the physico-chemical and entropy models and evaluate their ability to predict empirical measurements of HIV fitness. Definitions of all of our model parameters can be found in Table 1.

Modeling the HIV Fitness Landscape

In this study, we focus on two structurally similar models for describing a HIV fitness landscape: our physico-chemical based approach and Shannon entropy. Both models assume the fitness landscape is fixed and that each amino acid site affects viral fitness independent of the others. In the physico-chemical model the expected frequencies of the different amino acid residues are determined by their physico-chemical properties, the optimal residue for that site a^*, and the strength of consistent purifying selection G for a^*. These expected frequencies also depend on the physico-chemical weights $\vec{\theta}$, which are shared across a set of sites.

In contrast to our physico-chemical model, in the Shannon entropy model there are no shared parameters between sites, the optimal amino acid residue is assumed to be the most frequent one, and the frequency of the 19 other non-optimal amino acids residues are completely unconstrained and set to equal their observed frequencies for that site. Technically, the Shannon entropy model could be viewed as an unconstrained version of our physico-chemical model. In the Shannon entropy model there are 19 parameters per site, in our physico-chemical model there are only two, a^* and G.

A Physico-Chemical Model

For a given site, we assume the fitness w of amino acid residue a_j declines exponentially as a product of a site specific sensitivity parameter G_i and the distance d, in physico-chemical space, of a_j from the optimal amino acid residue for that site a^*. That is,

$$w\left(a_j \middle| a^*, G_i, \vec{\theta}\right) \propto \exp\left[-G_i d\left(a_j, a^* \middle| \vec{\theta}\right)\right]. \tag{1}$$

We use the euclidean distance function

$$d\left(a_j, a^* \middle| \vec{\theta}\right) = \sqrt{\theta_c(c_j - c_{a^*})^2 + \theta_l(l_j - l_{a^*})^2 + \theta_v(v_j - v_{a^*})^2}, \tag{2}$$

where c, p, and v represent the known composition, polarity, and molecular volume of a given amino acid residue, respectively. In contrast, the physico-chemical weighting terms $\vec{\theta} = (\theta_c, \theta_l, \theta_v)$ represent *a priori* unknown weights for each physico-chemical property. The physico-chemical weights $\vec{\theta}$ are are assumed to be shared across multiple sites and are estimated from the data. Our choice of physico-chemical properties follows Grantham (1974); other physico-chemical properties could be used instead (Sharma et al., 2013).

Following the example of other researchers in this field (Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2016) we treat HIV sequence data from different patients as independent samples from the evolutionary stationary distribution as described by Sella and Hirsh (2005). For now we ignore the effects of mutation bias and, as a result, the probability p of observing amino acid residue j, at site i is simply,

$$p_j\left(a_i^*, G_i, \vec{\theta}\right) = \frac{\exp\left[-G_i d\left(a_j, a_i^* \middle| \vec{\theta}\right) N_e\right]}{\sum_k \exp\left[-G_i d\left(a_k, a_i^* \middle| \vec{\theta}\right) N_e\right]} \tag{3}$$

where the composite parameter G_i' is the product of the site sensitivity and the effective population size N_e, i.e. $G_i' = G_i \times N_e$. Based on our model's assumptions and structure, G_i' represents a quantitative measure of the strength and efficacy of consistent purifying selection on a given site relative to genetic drift.

Given these assumptions, the probability of observing a set of amino acid residue

counts $\vec{x} = \{x_1, x_2, \ldots, x_{20}\}$ at a given site i follows a multinomial distribution. That is,

$$\Pr\left(\vec{x}\middle|a_i^*, G_i', \vec{\theta}\right) = \binom{x_T}{x_1\ x_2\ \ldots x_{20}} \prod_{j=1}^{20} p_j\left(a_i^*, G_i', \vec{\theta}\right)^{x_j} \tag{4}$$

$$= \mathcal{L}\left(a_i^*, G_i', \vec{\theta}\middle|\vec{x}\right) \tag{5}$$

where $x_T = \sum_j x_j$ is the total number of observations made at site i and $\mathcal{L}$ represents the likelihood function of the model parameters a_i^*, G_i', and $\vec{\theta}$ given the data $\vec{x}$. By maximizing Eq.(4) with respect to our model parameters, we can identify the most likely parameter values values and our confidence in them: the physico-chemical weighting terms $\vec{\theta}$, the optimal amino acid residue at each site a^*, and the N_e scaled sensitivity of HIV fitness to deviations in physico-chemical space for each site G'. By linking fitness to the physico-chemical of an amino acid residue, we effectively reduce the number of parameters in our multinomial model from 19 parameters per site to 2 parameters per site, G_i and a_i^* plus, depending on how we partition the data, 2 or 10 shared physico-chemical weight parameters $\vec{\theta}$. As a result, our physico-chemical model is highly 'unsaturated'.

The Entropy Model

While Shannon entropy is a measure of information of a set of messages, it is also proportional to the log of the probability of the data $\vec{x}$ under a multinomial model where the probability of each category p_j is equal to its observed relative frequency x_j/x_T where x_T is the number of observations. Setting $p_j = x_j/x_T$ not only makes intuitive sense, it is also equal to the maximum likelihood estimate (MLE) of p_j under a multinomial model. Thus Shannon entropy of a given site i with a set of counts $\vec{x}$ can be written as

$$E\left(\vec{x}\right) = -\sum_{j=1}^{20} x_j/x_T \ln\left(x_j/x_T\right) \tag{6}$$

and is related to the probability of the data $\vec{x}$ at the maximum likelihood values of its parameters $\vec{p}$ under a multinomial model

$$\Pr\left(\vec{x}\middle|\vec{p}=\frac{\vec{x}}{x_T}\right)=\binom{x_T}{x_1\ x_2\ \ldots x_{20}}\exp\left[-x_T E\left(\vec{x}\right)\right] \tag{7}$$

$$=\mathcal{L}\left(\vec{p}=\frac{\vec{x}}{x_T}\middle|\vec{x}\right) \tag{8}$$

The entropy of a site $E(\vec{x})$ can be interpreted in a number of different ways. One interpretation of $E(\vec{x})$ is as a diversity index for amino acid residues at a site or region (Jost, 2006). Another interpretation of $E(\vec{x})$ is that, under certain conditions, it is equal to the expected number of guesses one must make to infer the state of a site. A more statistically interpretation of $E(\vec{x})$ is as the average contribution of a category to a site's log-likelihood.

Despite having many potential interpretations, because of its descriptive and 'many to one' nature, there is no clear way of linking the Shannon entropy score of a site E and the strength of purifying selection for or against a particular amino acid. Finally, the inverse of $E(\vec{x})$ is used to describe the conservation of a site or region $1/E$ which has, in turn, been used as a heuristic measure of the strength of consistent purifying selection for the optimal amino acid residue at a site (Allen et al., 2005; Liu et al., 2012a). Because the entropy model has one p parameter per category and the sum of p must equal one, the entropy model is a saturated model with 19 free parameters per site.

Data

Amino Acid Sequences

To parameterize the models we used the Gag poly-protein of HIV subtype C MSA. We excluded linker regions and, as a result, analyzed 520 sites using 1058 curated sequences from the Los Alamos National Laboratory (LANL) database (Biophysics Group Los Alamos National Lab, 2016). This database contains filtered web alignments curated by the biophysics group at LANL. Details concerning the curation and filtering of the sequences in the alignment process can be found at `http://www.hiv.lanl.gov/`. Data was downloaded in September 2016. Sequences were processed to obtain counts for amino acid residue j at site i for each HXB2 site.

Empirical Estimates of HIV Fitness

In order to test whether our site sensitivity parameter G' is more biologically informative than the standard conservation metric $1/E$, we evaluated its ability to predict viral fitness using an *in vivo* data set and an *in vitro* data set. Specifically, we predict the fitness of genotypes that deviate from the optimal amino acid residue to decline with G'.

We used viral escape rates from a patient's CD8 T-cell response as a measure of *in vivo* fitness. Viral escape rates were calculated using Liu et al. (2012b) estimates of viral escape times for 24 reactive epitopes sites in 17 HIV-1 subtype B infected patients over 3 years. Escape time t was defined as the number of days between detection of the T-cell response and the time viral variants bearing that respective reactive epitope fell below 50%. These researchers found that that the mean conservation value $\overline{1/E}$ for an epitope is negatively correlated with escape time. Epitopes were defined as 8-11 amino acid long regions of the proteome recognized by patient specific T cell immune response. For consistency in the sign of our predictions for the *in vivo* and *in vitro* empirical data, we used we used escape rate $1/t$, rather than escape time t, as the predicted variable. The escape rate $1/t$ was compared to the average sensitivity $\bar{G}'$ and average conservation of an epitope $\overline{1/E}$.

We used viral replication rate in cell culture as a measure of *in vitro* HIV fitness. Rihn et al. (2013a) estimated the replicative fitness of 31 viruses bearing mutated capsid CA amino acid residues via spreading replication assay on human MT4 T-cell lines and peripheral blood monocytes. The CA mutants in this study were generated by creating a mutagenized CA library using a low fidelity PCR approach and then inserting the mutated CA sequences in replication competent proviral clones. Fitness was reported as % of the wild-type replication.

Model Parameterization

Because G' is always multiplied by our distance function d, there is an inherent lack of identifiability in our model. To solve this problem, $\vec{\theta}$ was constrained so that the sum of its values equaled 1. In order to identify a reasonable starting set of sensitivity values G' for the sites, we first optimized our physico-chemical model by fixing our

physico-chemical weights $\vec{\theta}$ to the values identified by Grantham (1974). We then parameterized our physico-chemical model using a two stage hill climbing optimization implemented using SCIPY packages (Walt et al., 2011). Briefly, for a given set of $\vec{\theta}$ values, we optimized G' and a^* for each site. We did so by finding the optimal G' value for each amino acid site using a sequential least squares programming iterative method SLSQP. The optimization for the G' values was initialized as a grid search using a vector containing a distribution of likely G' values.

We then chose the combination of a^* and G' with the largest log likelihood. We then used a constrained optimization by linear approximation COBYLA algorithm to optimize over $\vec{\theta}$ space, re-optimizing G' and a^* at each step as described above.

The optimization was first initialized at >50 random initial values in the physico-chemical weight space $\vec{\theta}$. In order to identify regions with high likelihood values in a computationally efficient manner, for each set of initial values we limited the number of rounds of optimization to 10-20 steps. The hill climbing optimization was reinitialized for the three sets of $\vec{\theta}$ with the highest likelihoods from the step limited optimization. The most likely parameter weight combination from these three starting points was then selected and assumed to represent the global optimum. In order to test whether the same $\vec{\theta}$ values were applicable across all protein regions, we compared our model fit with a single set of $\vec{\theta}$ values to one where $\vec{\theta}$ was allowed to vary between the protein regions p24, p17, p7, and p6.

Statistics

To avoid the effects of potential outliers, we excluded all hyper-conserved outlier positions, which we defined as sites with $1/E$ >100, from our statistical analyses. The distribution of these hyper-conserved sites values were clearly distinct from the remaining sites' $1/E$ and G' values (Fig. **??**).

Neither our measures of consistent purifying selection, site sensitivity G' and site conservation $1/E$, nor our empirical *in vivo* and *in vitro* fitness estimates were normally distributed. Therefore, we used Kendall's non-parametric rank correlation τ to measure the association between our consistent purifying selection metrics and the empirical data. We generated confidence intervals τ by bootstrapping the data 1000 times with

replacement. Statistical significance of τ was calculated via R package `Kendall` (McLeod, 2011).

In order to quantitatively describe and compare the site sensitivities G' of a given protein, we fitted three different probability distributions, the Log-Normal, Gamma, and Inverse-Gamma, to these values using maximum likelihood. We then used the likelihood ratio test to compare the null hypothesis H_0, the distribution of G' was similar across proteins, to the alternative hypothesis H_A, the distribution of G''s varied between protein. The H_A requires six more $\vec{\theta}$ parameters than H_0. We used R 3.3.2 and the package `bbmle` to estimate the maximum likelihood values and its likelihood profiling function to generate 95% confidence intervals for the probability distribution parameters describing a given set of G' values (R Core Team, 2017; Bolker and Team, 2016).

Results

Briefly, we find that our population genetics based site sensitivity terms G' are well correlated with the more commonly used entropy based conservation metric $1/E$. In terms of fitting the data, the entropy model, a parameter saturated model, has a substantially better AIC score than our physico-chemical model. This is not surprising given the fact that we have more than a 1000 observations for each site to estimate the entropy model's 19 site specific parameters. In contrast, we find that our physico-chemical model does a better job predicting empirical measurements of HIV fitness and can be used to test more refined hypotheses such as whether the nature of selection, as described by our physico-chemical weights $\vec{\theta}$ and G' values, varies between protein regions.

Variation in physico-chemical Weights $\vec{\theta}$

Although allowing the physico-chemical weights $\vec{\theta}$ to vary between Gag protein regions required the addition of just 6 additional $\vec{\theta}$ parameters, adding these parameters vastly improved the ability of our physico-chemical model to fit the sequence data by 5490 log-likelihood units (Table 2). Surprisingly, despite this vast improvement in model fit, the differences in $\vec{\theta}$ were actually quite small. These results clearly indicate that the effects of amino acid substitutions vary between proteins. Consequently, all of the

results we discuss below come from the model fit where $\vec{\theta}$ varies between proteins.

Sensitivity G' vs. Conservation $1/E$

Kendall's rank correlation τ between our model's site sensitivity parameter G' and the entropy model's site conservation $1/E$ metric indicate that they are well correlated with one another ($\tau = 0.567$, Fig 1 and Table 3). The degree of correlation between G' and $1/E$ varied substantially across Gag's different proteins (Fig 2). The nucleoprotein and matrix proteins were more strongly correlated than the capsid and nucleocapsid ($\tau =$ 0.791 and 0.625 vs. 0.456 and 0.344, respectively). G' and $1/E$ were less correlated in proteins possessing heterogeneous secondary structure (p17, p24, p7) than in p6, a disordered protein when part of the poly-peptide (Davey et al., 2014, Table 3).

Approximate Location of Fig 1

Approximate Location of Fig 2

Predicting Empirical Measurements of HIV Fitness

As expected, both site sensitivity G' and site conservation $1/E$ were negatively correlated with the *in vivo* measure of fitness escape rate ($\tau = -0.297$ and -0.230, respectively; Fig 3 and Table S1). However, only the correlation between G' and escape time was significant ($p = 0.046$ vs. 0.122 for $1/E$). Similarly, both G' and $1/E$ were negatively correlated with the effects of changes to the capsid protein *in vitro* replication fitness ($\tau = -0.381$ and -0.256, respectively ; Fig. 4 and Table S1). However, in this case τ of both G' and $1/E$ with *in vitro* fitness were significantly less than 0 ($p = 0.0021$ and 0.038, respectively).

Approximate Location of Fig 3

Approximate Location of Fig 4

Distribution of Site Sensitivities G'

We fitted three different distributions, the LogNormal, Gamma and Inverse-Gamma, to the physico-chemical model's site sensitivities G' for Gag's variable sites. Of the three distributions used, the LogNormal distribution performed the best for the entire poly-protein and the protein regions p6, p17, and p24 while the Gamma distribution

performed the best for the p7 region.(Tables S2 - S3). The null hypothesis H_0 that all G' come from the same distribution was rejected in favor of the alternative hypothesis H_A where the distribution of G' values varied between protein regions ($p < 0.XXX$; Table 5). The 95% Confidence intervals indicate that the LogNormal shape parameter σ did not vary significantly between p6, p17, and p24, but that location parameter μ was significantly lower in p6 and p17 than p24 (Fig. 5; Table S2).

Discussion

Although initially introduced as a metaphor for describing how evolution works, fitness landscapes have proven to be useful tools in theoretical and empirical research (for example, Kauffman, 1993; Wilke and Drummond, 2006; Gilchrist, 2007; Gilchrist et al., 2009; Wallace et al., 2013; Gilchrist et al., 2015, and citations below). Ideally, HIV fitness landscapes can help researchers predict when and where a virus will escape immune control (Barton et al., 2016) and develop effective vaccines (Ferguson et al., 2013; Shekhar et al., 2013). In a few cases HIV fitness landscapes have been estimated directly from experimental data (Hinkley et al., 2011; Kouyos et al., 2012; Mann et al., 2014; Rihn et al., 2013b), but in most studies the landscape is inferred from the ever growing libraries of genotype frequency data (Barton et al., 2016; Zanini et al., 2016; Mann et al., 2014; Ferguson et al., 2013; Deforche et al., 2008; Seifert et al., 2015).

In this study we utilize fundamental findings from the field of population genetics to infer HIV Gag's poly-protein physico-chemical fitness landscape using count data from the LANL HIV database. In contrast to explicitly mapping out HIV's fitness landscape, other researchers have used site conservation $1/E$, where E is the Shannon entropy of a site, as a proxy for the strength of consistent purifying selection (Rihn et al., 2013a; Ferrari et al., 2011; Liu et al., 2012b). While this interpretation may seem intuitive, $1/E$ is actually a summary statistic rather than a measure of the selection differential between a given pair of amino acid residues. Further, entropy based metrics such as $1/E$ result from fitting saturated models, where the number of parameters is on par with the number of data categories, to data. As a result, it is perhaps not surprising that the likelihood function for the entropy model fit the data better than our highly unsaturated physico-chemical model.

While saturated models have the advantage of being maximally flexible in terms of fitting data, this flexibility comes at the cost of being minimally informative about the processes generating the data. So it is perhaps also not surprising that our biologically grounded, physico-chemical model's site sensitivity G' did a better job than the entropy model's site conservation $1/E$ in predicting *in vivo* and *in vitro* HIV fitness measurements (Figs 3 and 4). These results suggest that our physico-chemical model is more efficient at extracting biological meaningful information from the sequence data we used to fit our models.

In addition to extracting more meaning from the data, because our physico-chemical model is derived from biological principles, it can be used to evaluate biologically based hypotheses. For example, we find that the impact of our three different physico-chemical traits $\vec{\theta}$ on fitness varies slightly, but significantly between protein regions (Table 2). These results indicates the effects of substituting one amino acid residue for another varies with its broader genetic context. This is, perhaps, unsurprising given that previous researchers have found that interior regions of the protein are more sensitive to changes in polarity (Nakai et al., 1988). Nevertheless, our ability to detect these slight regional differences in the character of the consistent purifying selection illustrates the statistical power of our physico-chemical modeling approach.

In addition, site sentivities G' values differ between protein regions in terms of the distribution that best describe the estimated values (LogNormal vs. Gamma). The fact that the G' values of different protein regions follow different distributions suggests that the contribution of each site to protein function varies between sites. For example, the LogNormal distribution suggests the G' values are the result many positive random independent variables acting in a multiplicative manner. In contrast, the Gamma distribution suggests that protein function depends on fluxing through different conformational states with exponential like waiting times for each state. Even amongst the three protein regions where site sensitivities G' are best described by a LogNormal distribution, they appear to come from distributions with different underlying parameters (Table **??**). For example, the G' values for region p24 were greater than p7 and p17, which is consistent with previous findings (Rihn et al., 2013a; Martinez-Picado et al., 2006). One testable hypothesis is that the differences between G' distributions result from the differences in secondary structure between protein region. This idea

could be tested by determining whether grouping sites by secondary structure (or some other feature such as distance from surface or active site) provides a better fit than grouping sites by protein region as we do here. Regardless of the underlying causes, we find that at a minimum the parameters underlying HIV's fitness landscape vary between protein regions.

While our physico-chemical model improves our ability to predict empirical HIV fitness data, there is still a substantial amount of noise in our predictions. This variation likely has a number of different sources. In terms of the data we are trying to predict, the *in vitro* spreading fitness measures suffer from the unnatural qualities of cell culture. Similarly, the *in vivo* epitope escape fitness measures likely includes substantial effects due to biological variation between patients, is baesed on a limited number of sample time points, and limited sequencing depth to determine genotype frequencies.

In terms of model shortcomings, there are many. For example, ur choice of physico-chemical properties to include in our distance function was based solely on Grantham's classic work (Grantham, 1974). Fortunately, testing the power of other physico-chemical properties is straightforward with our model. Further, in its current form, our physico-chemical model ignores the effect of mutation bias. Mutation bias, i.e. the fact that mutation rates between residues differ from one another, can also contribute to the probability of observing a particular codon and, in turn, amino acid. Although the effects of mutation bias likely to be unimportant at sites with large site sensitivities G', mutation bias can dominant the evolutionary outcome of sites under weak selection.

Moving on to shortcomings that are more challenging to over come, both the physico-chemical and entropy models assume statistical independence between patient samples. These assumptions also shared by the more complex Ising and Potts models and could be addressed by extending our approach to a phylogenetic framework (Beaulieu et al., In Review). Both the physico-chemical and entropy models also assume a single, invariant fitness peak centered around a site's optimal amino acid residue a^*. While it should be possible to extend our physico-chemical to allow for more than one peak in the amino acid residue landscape by modifying our distance function, we suspect our ability to reject the simpler hypothesis of a single peak would be weak, especially if they occurred in similar points in physico-chemical space. Allowing a^* to

explicitly switch over time, would be even more challenging than allowing for multiple fitness peaks. We expect detecting such switches would be very difficult without large, high quality and high resolution datasets.

Perhaps most glaringly shortcoming is the fact that our physico-chemical ignores epistasic effects. Epistatic interactions are likely ubiquitous and have been well documented in the virus literature (Koek et al., 2012; Brockman et al., 2007, 2010). Fortunately, we could extend our physico-chemical model to include the effects of epistatic interactions between sites in a similar manner to the Ising and Potts model. In a similar manner to how we use a physico-chemical distance function and site sensitivity G' to generate the predicted frequency of the 20 canonical amino acid residues, we could define a more complex distance function that would allow us to describe epistatic effects between sites and predict the probability of the 400 different possible site pairs of amino acid residues using a relatively small number of parameters. The end result would ideally be a more realistic model than the Ising model, but one that requires many fewer parameters, is more biologically informative, and is easier to fit than the Potts model.

In conclusion, we argue that one promising way to improve our ability to extract biologically meaningful information from sequence databases is to use well defined and biologically grounded models. Here we show how our physico-chemical model can improve our ability predict HIV fitness. Because our model is grounded in the field of population genetics, it is inherently well suited to describe evolutionary data. Because our model is well defined biologically, it provides a clear framework to test specific hypotheses about HIV protein sequence evolution and can serve as the basis for more complex, but parameter limited, models.

Tables

Table 1. Model parameters and variables contained in conservation calculations methods and sensitivity estimates.

Parameter	Definition
w_{ij}	Evolutionary fitness of amino acid residue j at site i.
N_e	Effective population size.
a_i^*	The optimal amino acid residue for site i.
G_i	Sensitivity of site i to deviation from a_i^*.
G_i'	Product of G_i and N_e.
$\vec{G'}$	vector of sensitivities scaled by N_e
$d(a_j, a^*)$	physico-chemical distance between residue j and a^*.
θ_c	Residue composition weighting term.
θ_l	Residue polarity weighting term.
θ_v	Residue molecular volume weighting term.
$\vec{\theta}$	Vector of physico-chemical residue weighting terms.
p_j	Probability of observing amino acid residue j at a given site.
x_j	Count of amino acid residue j plus 1 at a given site.
$\vec{x}$	Vector of amino acid residue counts at a given site.
$E(\vec{x})$	Shannon entropy for a set of residue counts $\vec{x}$ at a given site.
$1/E$	Measure of conservation of site as inverse of Shannon entropy.
$\mathcal{L}$	The likelihood of a set of parameters given the data.
L	The log-likelihood of a set of parameters given the data.

Table 2. Physico-Chemical Weights $\vec{\theta}$: Maximum likelihood values and corresponding log-likelihoods L for physico-chemical weights $\vec{\theta}$ individual protein regions p24, p17, p7, and p6 and for the entire Gag poly-protein. Parameters θ_c, θ_p, and θ_v represent the physico-chemical weights for amino acid residue composition, polarity, and molecular volume, respectively. ΔL is the difference between the L for the combined L for all four regions, p6, p7, p17, and p24, and the L for the Gag poly-protein and the Likelihood ratio test statistic for using single set of physico-chemical weights for Gag poly-protein vs. individual protein regions is $p < 10^{-3}$.

	Physico-Chemical Weights			
Protein	θ_c	θ_l	θ_v	L
Region				
p6	0.77299	0.22664	0.0003657	-11331.0
p7	0.77301	0.22613	0.0008545	-5086.3
p17	0.77329	0.22628	0.0000429	-26456.0
p24	0.77325	0.22634	0.0004128	-20679.0
			Total L	-63552
Poly-Protein	0.7738	0.22652	0.000478	-69040.
	Region and Poly-Protein ΔL			5488.

Table 3. Kendall correlation τ between G' and $1/E$ for each protein region along with their corresponding secondary structure. Secondary structure categorization based on (Davey et al., 2014) HIV Mutation Browser.

		Secondary Structure				
		Totals		Ordered Category		
Region	τ	Disordered	Ordered	Turn	Strand	Helix
p7	0.344	0.710	0.290	0.190	0.100	0.000
p24	0.456	0.320	0.680	0.092	0.021	0.570
p17	0.625	0.250	0.750	0.069	0.120	0.560
p6	0.791	1.000	0.000	0.000	0.000	0.000

Table 4. Comparison of Probability Distribution Fits To Site Sensitivity G' for Variable Sites Log-likelihood values for maximum likelihood fits of the Log-Normal, Gamma, and Inverse-Gamma distributions to the site sensitivities G' for the entire Gag poly-peptide and individual protein regions. Region L is the sum of L for the 4 protein regions when fitted using the same distribution function.

	Log-Likelihood L		
Protein	Gamma	Inverse Gamma	Log-Normal
Region			
p17	−340.249	−345.387	−337.849
p24	−504.851	−510.870	−502.455
p6	−118.841	−117.708	−116.457
p7	−115.864	−123.631	−118.391
Total L	−1079.805	−1097.596	−1075.152
Poly-Peptide	−1103.571	−1131.988	−1102.144

Table 5. Model Selection for Common vs. Region Specific Distributions of Site Sensitivities G' Comparison of null model H_0 where the distribution of physico-chemical site sensitivities G' follow a single, Log-Normal distribution vs. an alternative model H_A where the distribution of G''s vary between protein region, following a Gamma distribution for p7 and a separate Log-Normal for each of the remaining regions as presented in Table 4.

Model	Number of Parameters	L	AIC
Region	8	-1072.626	-2161.252
Poly-protein	2	-1102.144	-2208.288
		ΔAIC	47.036

Figures

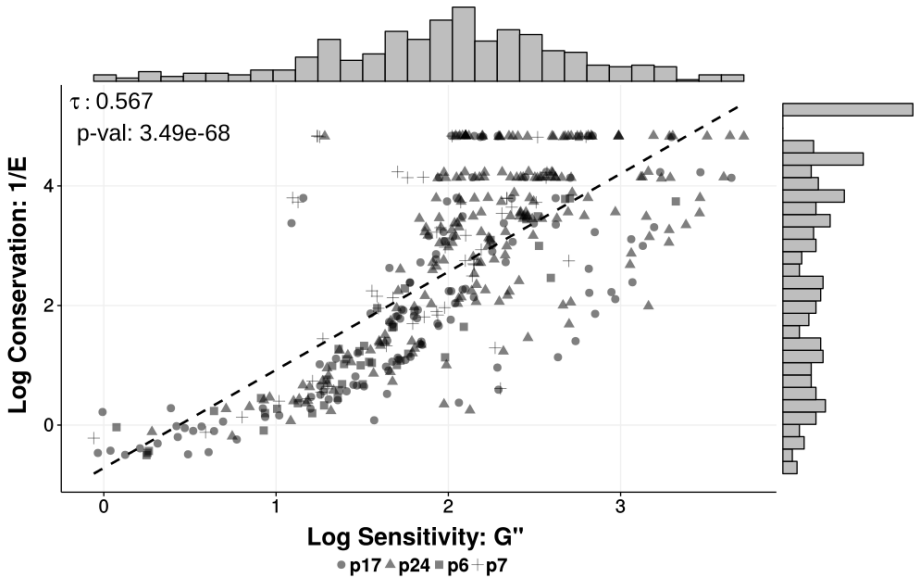

Fig 1. Rank correlation of each position's estimated physico-chemical sensitivity G' with its conservation value $1/E$. Positions total 520 in the Gag poly-protein excluding spacer regions. Estimates of sensitivity G' and conservation $1/E$ estimated and derived from 1000 sample MSA of the Gag poly-protein for HIV-1 subtype C. Sensitivity and conservation concord with each other with a significant Kendall τ of 0.567.

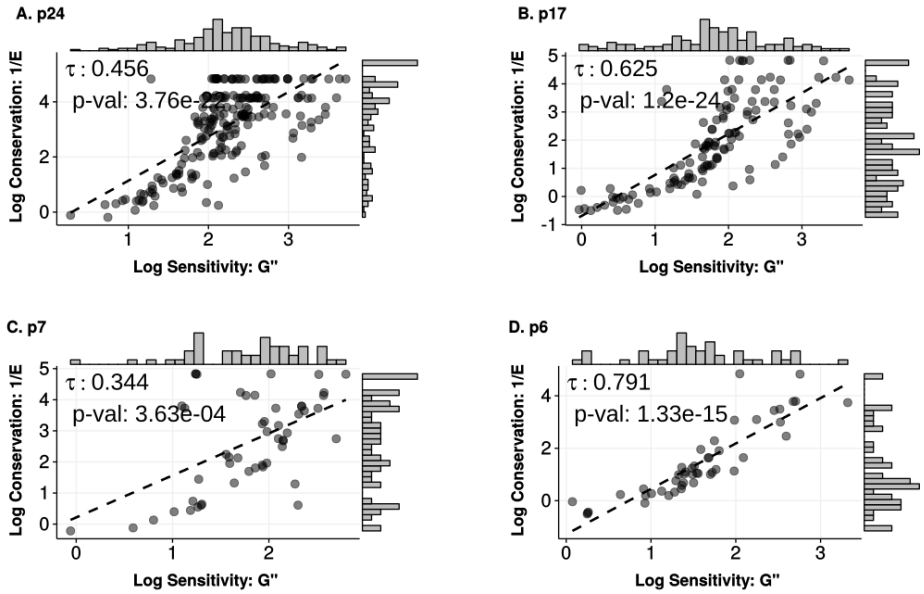

Fig 2. Kendall τ correlation of site sensitivity G' and site conservation $1/E$ in the four protein regions in the Gag poly-protein A) capsid p24 $\tau = 0.456$, B) matrix p17 $\tau = 0.625$, C) nucleo-capsid p7 $\tau = 0.344$, and D) nucleo-protein $\tau = 0.791$.

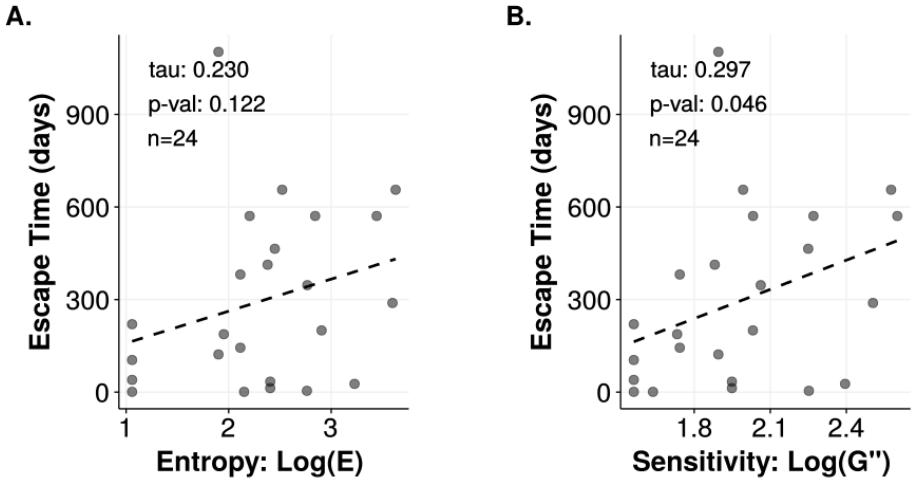

Fig 3. Metrics vs Escape Time Correlation of entropy and sensitivity with the escape time of 24 epitopes in the Gag poly-protein A) Correlation of the escape time of an epitope in days with the epitope's entropy B) Correlation of the escape time of an epitope in days with the epitope's physico-chemical sensitivity.

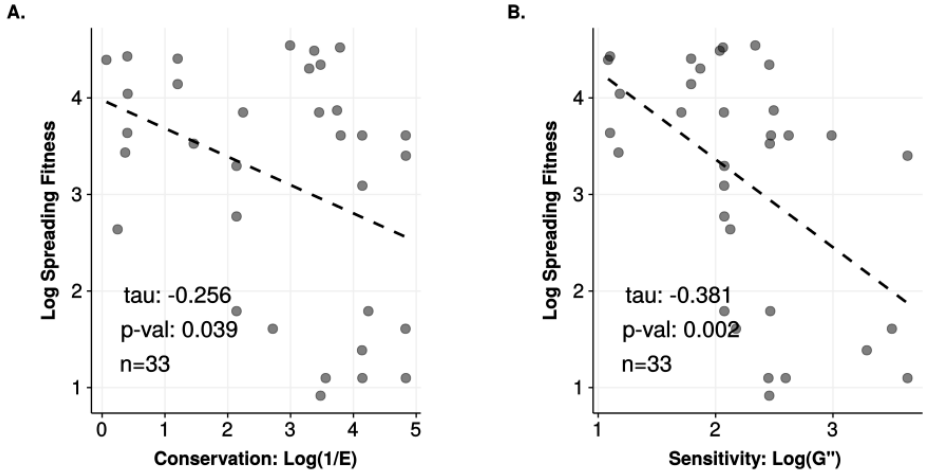

Fig 4. Metrics vs Spreading fitness Correlation of conservation and sensitivity of residues with spreading fitness of 31 viral strains bearing mutations in the corresponding residues A) Correlation of the assayed spreading fitness of the mutated virus in days with the correspondingly position's entropy B) Correlation of the assayed spreading fitness of the mutated virus in days with the corresponding position's physico-chemical sensitivity.

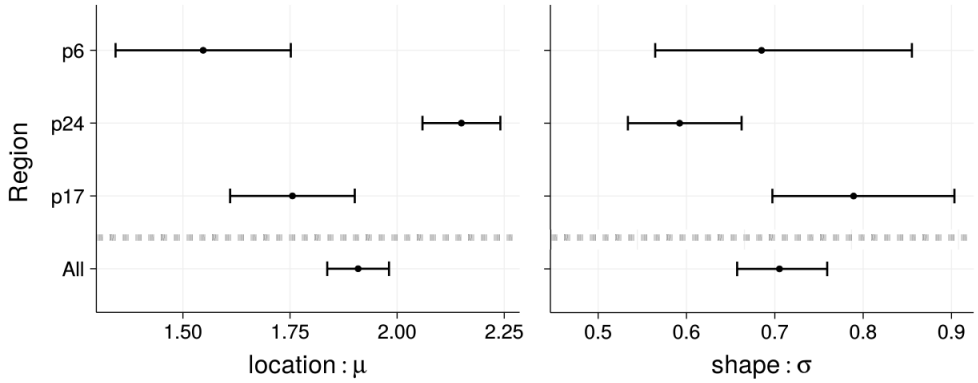

Fig 5. MLE of Log-Normal parameters for each region Panels display Log-Normal parameter fits for each region and all regions collectively. **A.** The location parameter for the capsid's (p24) distribution is statically higher compared to the other regions and the nucleo-capsid's (p7) distribution is statistically lower than the poly-protein's (All) distribution. **B.** Shows location parameter σ for each region with 95% confidence intervals estimated by likelihood profiling. The capsid (p24) and matrix (p17) distributions have a shape parameters distinct from each-other

References

Acevedo, A., L. Brodsky, and R. Andino. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686–690.

Allen, T. M., M. Altfeld, S. C. Geer, E. T. Kalife, C. Moore, et al. 2005. Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. Journal of Virology 79:13239–13249.

Autran, B., R. L. Murphy, D. Costagliola, R. Tubiana, B. Clotet, et al. 2008. Greater viral rebound and reduced time to resume antiretroviral therapy after therapeutic immunization with the ALVAC-HIV vaccine (vCP1452). AIDS (London, England) 22:1313–1322.

Barton, J. P., N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, et al. 2016. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nature Communications 7:11660.

Beaulieu, J., B. O'Meara, C. Landerer, J. J. Chai, and M. A. Gilchrist. In Review. Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: A nested modeling approach. Systematic Biology .

Berg, J. and M. Lässig. 2003. Stochastic evolution and transcription factor binding sites. Biophysics 48:S36–S44.

Berg, J., S. Willmann, and M. Lassig. 2004. Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology 4:1–12.

Biophysics Group Los Alamos National Lab. 2016. Sequence Alignments.

Bolker, B. and R. D. C. Team. 2016. bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.18.

Brockman, M. A., Z. L. Brumme, C. J. Brumme, T. Miura, J. Sela, et al. 2010. Early Selection in Gag by Protective HLA Alleles Contributes to Reduced HIV-1 Replication Capacity That May Be Largely Compensated for in Chronic Infection. Journal of Virology 84:11937–11949.

Brockman, M. A., A. Schneidewind, M. Lahaie, A. Schmidt, T. Miura, et al. 2007. Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A. Journal of Virology 81:12608–12618.

Calcott, B. 2008. Assessing the fitness landscape revolution. Biology & Philosophy 23:639–657.

Charlesworth, B. and S. Rouhani. 1988. The probability of peak shifts in a founder population .2. an additive polygenic trait. Evolution 42:1129–1145.

Cuevas, J. M., R. Geller, R. Garijo, J. Lpez-Aldeguer, and R. Sanjun. 2015. Extremely High Mutation Rate of HIV-1 In Vivo. PLoS Biology 13.

Davey, N. E., V. P. Satagopam, S. Santiago-Mozos, C. Villacorta-Martin, T. A. M. Bharat, et al. 2014. The hiv mutation browser: A resource for human immunodeficiency virus mutagenesis and polymorphism data. Plos Computational Biology 10:e100395.

Deforche, K., R. Camacho, K. V. Laethem, P. Lemey, A. Rambaut, et al. 2008. Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. Bioinformatics 24:34–41.

Dietrich, M. R. and R. Skipper. 2012. A Shifting Terrain: A Brief History of the Adaptive Landscape. The adaptive landscape in evolutionary biology :3–15.

Draghi, J. A., T. L. Parsons, G. P. Wagner, and J. B. Plotkin. 2010. Mutational robustness can facilitate adaptation. Nature 463:353–355.

Ferguson, A. L., J. K. Mann, S. Omarjee, T. Ndungu, B. D. Walker, et al. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity 38:606–617.

Ferrari, G., B. Korber, N. Goonetilleke, M. K. P. Liu, E. L. Turnbull, et al. 2011. Relationship between Functional Profile of HIV-1 Specific CD8 T Cells and Epitope Variability with the Selection of Escape Mutants in Acute HIV-1 Infection. PLOS Pathog 7:e1001273.

Fontana, W. 2002. Modelling 'evo-devo' with RNA. Bioessays 24:1164–1177.

Gavrilets, S. 1997. Evolution and speciation on holey adaptive landscapes. Trends In Ecology & Evolution 12:307–312.

Gavrilets, S. 2004. Fitness Landscapes and the Origin of Species. No. 41 in Monographs in Population Biology, Princeton University Press, Princeton, NJ.

Gilchrist, M., P. Shah, and R. Zaretzki. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183:1493–1505.

Gilchrist, M. A. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. Molecular Biology and Evolution 24:2362–2373.

Gilchrist, M. A., W.-C. Chen, P. Shah, C. L. Landerer, and R. Zaretzki. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. Genome Biology and Evolution 7:1559–1579.

Goulder, P. J. R. and D. I. Watkins. 2004. HIV and SIV CTL escape: implications for vaccine design. Nature Reviews Immunology 4:630–640.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, et al. 2011. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nature Genetics 43:487–489.

Iwasa, Y. 1988. Free fitness that always increases in evolution. Journal of Theoretical Biology 135:265–281.

Johnston, M. I. and A. S. Fauci. 2008. An HIV Vaccine Challenges and Prospects. New England Journal of Medicine 359:888–890.

Jost, L. 2006. Entropy and diversity. Oikos 113:363–375.

Kauffman, S. and S. Levin. 1987. Towards a general-theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology 128:11–45.

Kauffman, S. A. 1993. Origins of Order. Oxford University Press.

Kouyos, R. D., G. E. Leventhal, T. Hinkley, M. Haddad, J. M. Whitcomb, et al. 2012. Exploring the Complexity of the HIV-1 Fitness Landscape. PLoS Genet 8:e1002551.

Koek, M., S. Henke, K. G. akov, G. B. Jacobs, A. Schuch, et al. 2012. Mutations in HIV-1 gag and pol Compensate for the Loss of Viral Fitness Caused by a Highly Mutated Protease. Antimicrobial Agents and Chemotherapy 56:4320–4330.

Lande, R. 1985. Expected time for random genetic drift of a population between stable phenotypic states. Proceedings Of The National Academy Of Sciences Of The United States Of America 82:7641–7645.

Lande, R. 1986. The dynamics of peak shifts and the pattern of morphological evolution. Paleobiology 12:343–354.

Lande, R. and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lässig, M. 2007. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 8 Suppl 6:S7.

Lemey, P., A. Rambaut, and O. G. Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS reviews 8:125–140.

Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, et al. 2012a. Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. Journal of Clinical Investigation .

Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, et al. 2012b. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. Journal of Clinical Investigation .

Liu, Y., J. McNevin, J. Cao, H. Zhao, I. Genowati, et al. 2006. Selection on the Human Immunodeficiency Virus Type 1 Proteome following Primary Infection. Journal of Virology 80:9519–9529.

Lorenzo-Redondo, R., S. Delgado, F. Morn, and C. Lopez-Galindez. 2014. Realistic Three Dimensional Fitness Landscapes Generated by Self Organizing Maps for the Analysis of Experimental HIV-1 Evolution. PLoS ONE 9:e88579.

Mann, J. K., J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, et al. 2014. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. PLoS Comput Biol 10:e1003776.

Martinez-Picado, J., J. G. Prado, E. E. Fry, K. Pfafferott, A. Leslie, et al. 2006. Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. Journal of virology 80:3617–3623.

McCandlish, D. M., C. L. Epstein, and J. B. Plotkin. 2015. Formal properties of the probability of fixation: Identities, inequalities and approximations. Theoretical Population Biology 99:98–113.

McLeod, A. 2011. Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2.

Moradigaravand, D., R. Kouyos, T. Hinkley, M. Haddad, C. J. Petropoulos, et al. 2014. Recombination Accelerates Adaptation on a Large-Scale Empirical Fitness Landscape in HIV-1. PLoS Genet 10:e1004439.

Mustonen, V. and M. Lässig. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet 25:111–119.

Nakai, K., A. Kidera, and M. Kanehisa. 1988. Cluster-analysis of amino-acid indexes for prediction of protein-structure and function. Protein Engineering 2:93–100.

Niklas, K. J. 1994. Morphological evolution through complex domains of fitness. Proceedings of the National Academy of Sciences of the United States of America 91:6772–6779.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rihn, S. J., S. J. Wilson, N. J. Loman, M. Alim, S. E. Bakker, et al. 2013a. Extreme Genetic Fragility of the HIV-1 Capsid. PLoS Pathog 9:e1003461.

Rihn, S. J., S. J. Wilson, N. J. Loman, M. Alim, S. E. Bakker, et al. 2013b. Extreme genetic fragility of the hiv-1 capsid. PLoS Pathogens :e1003461.

Rolland, M., S. Manocheewa, J. V. Swain, E. C. Lanxon-Cookson, M. Kim, et al. 2013. HIV-1 Conserved-Element Vaccines: Relationship between Sequence Conservation and Replicative Capacity. Journal of Virology 87:5461–5467.

Salemi, M. 2013. The Intra-Host Evolutionary and Population Dynamics of Human Immunodeficiency Virus Type 1: A Phylogenetic Perspective. Infectious Disease Reports 5.

Seifert, D., F. D. Giallonardo, K. J. Metzner, H. F. Gnthard, and N. Beerenwinkel. 2015. A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory. Genetics 199:191–203.

Sella, G. and A. E. Hirsh. 2005. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America 102:9541–9546.

Sharma, A., K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, et al. 2013. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinformatics 14:233.

Shekhar, K., C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, et al. 2013. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 88:062705.

Wallace, E. W. J., E. M. Airoldi, and D. A. Drummond. 2013. Estimating selection on synonymous codon usage from noisy experimental data. Molecular Biology And Evolution 30:1438–1453.

Walt, S. v. d., S. C. Colbert, and G. Varoquaux. 2011. The NumPy Array: A Structure for Efficient Numerical Computation.

Wilke, C. O. and D. A. Drummond. 2006. Population genetics of translational robustness. Genetics 173:473–481.

Wilke, C. O. and D. A. Drummond. 2010. Signatures of protein biophysics in coding sequence evolution. Current Opinion In Structural Biology 20:385–389.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *In* D. F. Jones, ed., Proceedings of the Sixth International Congress on Genetics, vol. 1. Austin, TX.

Wright, S. 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies., vol. 2. University of Chicago Press.

Zanini, F., V. Puller, J. Brodin, J. Albert, and R. Neher. 2016. In-vivo mutation rates and fitness landscape of HIV-1. bioRxiv :045039.

Part I

Appendix

Supporting Materials

Statistics

Tables

Table S1. Model Predictions of HIV Fitness: Kendall τ correlations of site sensitivity G' and site conservation $1/E$ with *in vivo* and *in vitro* measures of HIV fitness with bootstrapped 95% confidence intervals.

	Data Type	
Parameter	*in vivo*	*in vitro*
Physico-Chemical G'	-0.297 (-0.008, -0.560)	-0.381 (-0.448, -0.044)
Entropy $1/E$	-0.230 (0.074, -0.517)	-0.256 (-0.545, -0.212)

Table S2. Log-Normal Fits to Physico-Chemical Site Sensitivities G' Maximum likelihood parameters and corresponding 95% confidence intervals for the Gamma distribution fitted to the G' for the entire polypeptide and the protein regions p6, p7, p17, p24. 95% confidence intervals calculated from the parameter's log-likelihood L profile.

Region	L	μ	σ
p7	-118.39	1.779 (1.609, 1.948)	0.567 (0.468, 0.708)
p6	-116.46	1.547 (1.343, 1.752)	0.685 (0.565, 0.855)
p17	-337.85	1.756 (1.610, 1.901)	0.789 (0.697, 0.903)
p24	-502.46	2.150 (2.059, 2.241)	0.592 (0.534, 0.662)
Poly-Protein	-1102.14	1.909 (1.837, 1.981)	0.705 (0.657, 0.759)

Table S3. Gamma Distribution Fits to Physico-Chemical Site Sensitivities G' Maximum log-likelihood L under the given maximum likelihood parameters estimates and their corresponding 95% confidence intervals for the Gamma distribution. Distribution fitted to the estimated site sensitivities G' for the entire polypeptide and separate protein regions: p6, p7, p17, and p24. 95% confidence intervals calculated from the parameter's likelihood profile.

Region	L	α	β
p7	-118.39	1.779 (1.609, 1.948)	0.567 (0.468, 0.708)
p6	-116.46	1.547 (1.343, 1.752)	0.685 (0.565, 0.855)
p17	-337.85	1.756 (1.610, 1.901)	0.789 (0.697, 0.903)
p24	-502.46	2.150 (2.059, 2.241)	0.592 (0.534, 0.662)
Poly-Protein	-1102.14	1.909 (1.837, 1.981)	0.705 (0.657, 0.759)