A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory

David Seifert,*,† Francesca Di Giallonardo,‡ Karin J. Metzner,‡
Huldrych F. Günthard,‡ and Niko Beerenwinkel*,†,1
*Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland, †Swiss Institute of Bioinformatics,
Basel 4058, Switzerland, and ‡Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich,
University of Zurich, Zurich 8091, Switzerland

ABSTRACT Fitness is a central quantity in evolutionary models of viruses. However, it remains difficult to determine viral fitness experimentally, and existing *in vitro* assays can be poor predictors of *in vivo* fitness of viral populations within their hosts. Next-generation sequencing can nowadays provide snapshots of evolving virus populations, and these data offer new opportunities for inferring viral fitness. Using the equilibrium distribution of the quasispecies model, an established model of intrahost viral evolution, we linked fitness parameters to the composition of the virus population, which can be estimated by next-generation sequencing. For inference, we developed a Bayesian Markov chain Monte Carlo method to sample from the posterior distribution of fitness values. The sampler can overcome situations where no maximum-likelihood estimator exists, and it can adaptively learn the posterior distribution of highly correlated fitness landscapes without prior knowledge of their shape. We tested our approach on simulated data and applied it to clinical human immunodeficiency virus 1 samples to estimate their fitness landscapes *in vivo*. The posterior fitness distributions allowed for differentiating viral haplotypes from each other, for determining neutral haplotype networks, in which no haplotype is more or less credibly fit than any other, and for detecting epistasis in fitness landscapes. Our implemented approach, called *QuasiFit*, is available at http://www.cbg.ethz.ch/software/quasifit.

FITNESS is a central quantity in evolutionary biology. It can be regarded as a measure of reproductive capacity of each individual. In evolving populations, individuals with higher fitness can outcompete those with lower fitness. Fitness depends on the genetic composition of the individual's haplotype, *i.e.*, the allelic constellation of multiple loci of its genome, and on the environment, *i.e.*, the host conditions for viral reproduction. Determining the fitness of viruses in a population is experimentally difficult and laborious, because individual virus particles need to be isolated and analyzed separately. *In vitro* determination of viral fitness usually involves enzymatic, growth competition, or monoinfection assays (Quiñones-Mateu and Arts 2002). A drawback to all *in vitro* measurements of viral fitness is the removal of viruses from the environment to which they have adapted. Such estimates disregard effects on the fitness landscape deriving from the natural *in vivo* environment.

RNA viruses, such as human immunodeficiency virus (HIV), have very high mutation rates (Rezende and Prasad 2004), and the number of haplotypes that arise in the normal course of intrahost evolution can be extremely large (Steinhauer and Holland 1987). Thus, HIV populations change and explore sequence space on timescales that are much shorter than those of higher eukaryotes. As such, RNA viruses lend themselves to being studied as model systems of evolutionary theory. In clinical settings, viral fitness is also of great interest. Disease progression, the formation of escape mutants, and ultimately treatment failure depend on viral fitness (Clavel and Hance 2004; Beerenwinkel *et al.* 2013).

The frequencies of viral haplotypes are determined by evolutionary parameters, including mutation rate and fitness. With the advent and exponential decrease in cost of next-generation sequencing (NGS) data (Niedringhaus *et al.*

Copyright © 2015 by the Genetics Society of America
doi: 10.1534/genetics.114.172312
Manuscript received July 31, 2014; accepted for publication November 10, 2014; published Early Online November 17, 2014.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.172312/-/DC1.
[1] Corresponding author: Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology, Mattenstrasse 26, 4058 Basel, Switzerland. E-mail: niko.beerenwinkel@bsse.ethz.ch

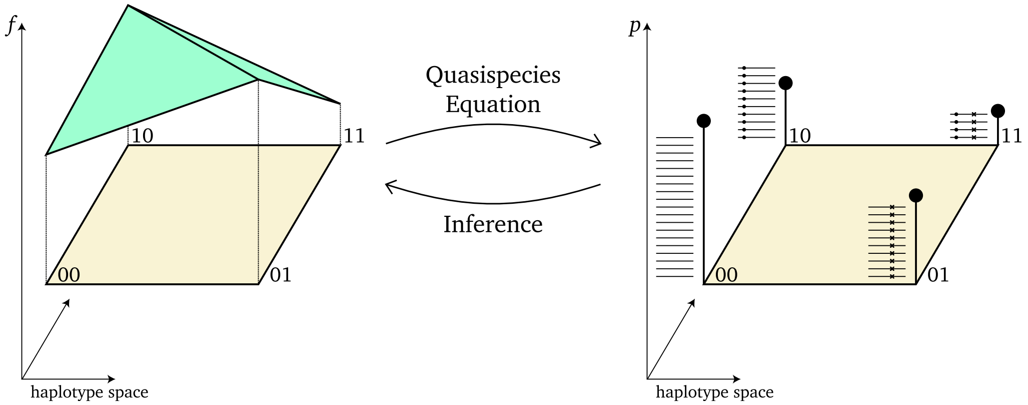

Figure 1 Schematic illustration of fitness landscapes and quasispecies. On the left is a fitness landscape on a simple biallelic two-locus genome. Here, 0 indicates a wild-type allele and 1 indicates a mutant allele and the vertical axis indicates fitness *f*. The fitness landscape also includes epistasis, as the fitness of the double mutant is not additive in the main fitness effects of the two mutant alleles. For this fitness landscape, at equilibrium, the quasispecies equation yields the quasispecies, *i.e.*, the mutation–selection equilibrium (right-hand side). The vertical bars indicate the relative frequency *p* of each haplotype. In practice, these frequencies are not known but can be estimated from next-generation sequencing data. The stacked horizontal lines represent reads from a sequencing experiment and amount to a finite sample of the quasispecies. Solid circles and crosses indicate mutant alleles at the two loci. Due to the sampling variance inherent in the finite sample, the number of reads will not match the frequencies exactly. Note that fitness values for the haplotypes need not show a strong linear correlation with the haplotype frequencies, as mutational coupling can obscure this relationship. Given a finite sample of the population, we aim to infer properties of the fitness landscape (right-to-left arrow).

2011), the technical prerequisites for affordable in-depth personalized diagnostics are within reach. Acevedo *et al.* (2014) have shown that inference of marginal fitness effects of single nucleotide variants in viruses with NGS is possible already today. High-quality NGS data for intrahost viral populations will become ubiquitous in the near future, making *in vivo* fitness analysis on the basis of such data possible.

A fitness landscape is the association of a real, non-negative fitness value to each haplotype. Fitness landscapes can be perfectly correlated, "Mount Fuji"-like with strong correlation between the fitnesses of closely related haplotypes or, on the other end of the spectrum, extremely rugged and spiky ("house of cards"), with no correlation of fitness values between related haplotypes (Gavrilets 2004). The effect of multiple alleles acting in concert to confer a fitness unexpected from the individual alleles is termed epistasis. The key factor for the ruggedness of a fitness landscape is the degree of epistasis involved in shaping it.

Several computational methods have been proposed for predicting *in vitro* fitness from viral sequence (Segal *et al.* 2004; Deforche *et al.* 2008; Ma *et al.* 2010; Hinkley *et al.* 2011; Ferguson *et al.* 2013) and for analyzing the structure of HIV *in vitro* fitness landscapes (Beerenwinkel *et al.* 2007a, b; Kouyos *et al.* 2012). For example, Hinkley *et al.* (2011) have performed large-scale *in vitro* fitness estimation of the HIV-1 protease and reverse transcriptase in the absence of drugs as well as in the presence of 15 antiretroviral drugs. However, neither this nor any other published study considers intrahost viral genetic diversity and hence none can account for *in vivo* fitness effects deriving from the host environment.

To estimate *in vivo* fitness landscapes and without direct observation of the growth kinetics of the viral population, an evolutionary model is required that links fitness to haplotype frequencies. Here we employ, for this purpose, the quasispecies model, an established model of intrahost viral evolution (Eigen and Schuster 1977; Burch and Chao 2000; Vignuzzi *et al.* 2005; Metzner *et al.* 2009; Domingo *et al.* 2012), which is mathematically tractable. One of the predictions of quasispecies theory that has stood the test of time is the existence of an error threshold in viral replication. If the mutation rate of a virus lies above this critical threshold, then mutation will cause genetic information to be lost. Anderson *et al.* (2004) have shown this phenomenon to exist in practice with the use of mutagenic nucleosides.

In this article, we establish a computational framework for estimating fitness from NGS data based on quasispecies theory. The Markov chain Monte Carlo (MCMC) sampler developed here infers the posterior distribution of fitness landscapes given NGS count data obtained from mixed intrahost virus populations (Figure 1). This inference scheme makes use of cross-sectional data, which are common in clinical settings, and where time series data are scarce.

Methods

The quasispecies model

The quasispecies model describes the evolution of an infinite population of DNA (or RNA) sequences (Eigen and Schuster 1977). We define the DNA alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and DNA sequence space of a genomic region of length L as the Cartesian product $\mathcal{A}^L = \{(a_1, ..., a_L) \mid a_i \in \mathcal{A}\}$. The elements of $\mathcal{A}^L$ are synonymously referred to as viral haplotypes, genotypes, or strains. They are indexed by $i = 1, \ldots, m = |\mathcal{A}|^L$.

The quasispecies equation is a first-order coupled nonlinear differential equation describing the temporal dynamics of a population subject to mutation and selection. Selection results from increased replication due to higher fitness, and coupling between haplotypes is maintained by mutation. A quasispecies is a cloud of closely connected haplotypes that evolve according to

$$\dot{p}_i(t) = \sum_{j=1}^{m} p_j(t) f_j q_{ji} - p_i(t) \sum_{j=1}^{m} p_j(t) f_j, \quad i \in \{1, \dots, m\}, \tag{1}$$

where $p_i(t)$ denotes the relative frequency at time $t \geq 0$, f_i the fitness of haplotype i, and q_{ji} the probability of haplotype j mutating into haplotype i upon replication. The term $p_j(t) f_j q_{ji}$ in the first sum denotes the flux, *i.e.*, the approximate amount of haplotype j mutating into haplotype i per unit time. The second sum is a normalization constant and ensures that the frequencies of all haplotypes sum to 1 for all time points. The fitness landscape $\mathbf{f} = (f_1, \dots, f_m)^T \in \mathbb{R}_+^m$ is static and does not change with time. In matrix notation, the quasispecies Equation 1 becomes

$$\dot{\mathbf{p}}(t) = \mathbf{Q}^T \mathrm{diag}(\mathbf{f}) \mathbf{p}(t) - \phi(\mathbf{p}(t), \mathbf{f}) \mathbf{p}(t) \tag{2}$$

with $\mathbf{p} = (p_1, \dots, p_m)^T \in \Delta^{m-1} = \{(x_1, \dots, x_m)^T \in \mathbb{R}_+^m \mid \sum_{i=1}^m x_i = 1\}$, the $(m-1)$-dimensional probability simplex, average fitness $\phi(\mathbf{p}(t), \mathbf{f}) = \mathbf{p} \cdot \mathbf{f}$, and mutation probability matrix $\mathbf{Q} = (q_{ij})$. In the literature, the mutation–selection matrix $\mathbf{Q}^T \mathrm{diag}(\mathbf{f})$ is often denoted $\mathbf{W}$ (Eigen *et al.* 1988; Wilke 2005).

Equilibrium distribution: The equilibrium distribution $\mathbf{p}^*$ of the quasispecies Equation 2 is well known (Eigen *et al.* 1988). It is obtained by setting $\dot{\mathbf{p}}(t) = 0$, such that

$$\phi(\mathbf{p}, \mathbf{f}) \mathbf{p}^* = \mathbf{Q}^T \mathrm{diag}(\mathbf{f}) \mathbf{p}^*. \tag{3}$$

If $\mathbf{Q}$ has only positive entries, then every haplotype has a nonzero probability of mutating into any other haplotype, and the transition matrix $\mathbf{Q}$ is irreducible and a Perron matrix. If all fitness values are also positive, then the matrix $\mathbf{Q}^T \mathrm{diag}(\mathbf{f})$, which is a column-wise reweighting of $\mathbf{Q}^T$, is still a Perron matrix. As a consequence of the Perron–Frobenius theorem, there exists a unique real eigenvalue ϕ larger than the absolute value of the real part of any other eigenvalue. To determine the equilibrium distribution (Equation 3), we first calculate the largest real eigenvalue ϕ of $\mathbf{Q}^T \mathrm{diag}(\mathbf{f})$ and its associated eigenvector possessing only positive components. Normalizing this eigenvector by dividing it by the sum of its components yields the global equilibrium distribution of the quasispecies equation (Eigen *et al.* 1988; Nowak and May 2000).

The mutation–selection equilibrium $\mathbf{p}^*$ is referred to as the quasispecies in quasispecies theory. The quasispecies model can be formulated for finite populations, where it is very similar to the Wright–Fisher model (Park *et al.* 2010; Musso 2012). When the effective population size N_e and the mutation rate μ are such that $N_e \mu > 1$, then the quasispecies is closely related to classical mutation–selection equilibrium from population genetics as exemplified in Wright's equation (Wilke 2005).

The single globally stable mutation–selection equilibrium is one appealing feature of the quasispecies equation, making it more tractable than other models of evolution. For fixed $\mathbf{Q}$, we denote by $\mathcal{Q}^{m-1} \subsetneq \Delta^{m-1}$ the set of all stationary distributions $\mathbf{p}^*$ arising under the quasispecies model for positive fitness landscapes $\mathbf{f} \in \mathbb{R}_+^m$.

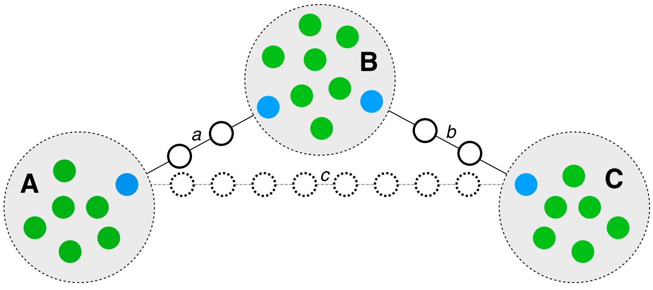

Figure 2 Illustration of the procedure to make the haplotype graph G_1 connected. Solid circles indicate observed haplotypes from sequencing in sequence space. Observed haplotypes are grouped in three connected components denoted with uppercase letters in boldface type, **A**, **B**, and **C**. The shortest mutational path between any pair of connected components is indicated with lines and labeled with lowercase italic letters *a*, *b*, and *c*. Blue solid circles indicate those haplotypes forming the endpoints of the shortest path between any pair of connected components. The aim is to make the whole graph connected while inserting the least number of unobserved haplotypes. In this case, the two paths *a* and *b* require inserting only two unobserved haplotypes each, while path *c* requires inserting eight haplotypes and is not used. The edges *a* and *b* form the minimum spanning tree of connected components and the haplotypes representing the solid circles are inserted.

Haplotype space and mutation probabilities: Working with the full combinatorial DNA sequence space $\mathcal{A}^L$ is infeasible, because its dimension grows exponentially in L. To employ this model for real data, we work on a reduced haplotype subset $\mathcal{H} \subset \mathcal{A}^L$ that is sufficiently small to allow for computational analysis but large enough to account for HIV's large heterogeneity. In practice, the haplotype space $\mathcal{H}$ contains all haplotypes observed in the sequencing data, plus additional unobserved ones, such that it is sufficiently connected, as detailed below. Working on a reduced haplotype space also reflects biological reality, where most DNA sequences of length L do not encode viable viruses and hence are extremely unlikely to arise in the course of HIV evolution and can safely be ignored.

To define the mutation probabilities (supporting information, File S1, section 1), we assume an identical per-site mutation probability $\mu > 0$. This constant reflects the fidelity of reverse transcription and is ~3 × 10^{-5} per replication for HIV. We denote by $d(i, j)$ the Hamming distance between haplotypes i and j, *i.e.*, the number of loci at which they differ. For a haplotype subset $\mathcal{H}$ of cardinality n, we define $\mathbf{Q} = (q_{ij})$ by setting

$$q_{ij} = \left(\frac{\mu}{|\mathcal{A}| - 1} \right)^{d(i,j)} \cdot (1 - \mu)^{L - d(i,j)} \tag{4}$$

for all $i \neq j$, and $q_{ii} = 1 - \sum_{j \in \{1, \dots, n\} \setminus \{i\}} q_{ij}$, for all $i = 1, \dots, n$. In the case that the uniform mutation probabilities are considered too restricted, a more general two-rate model

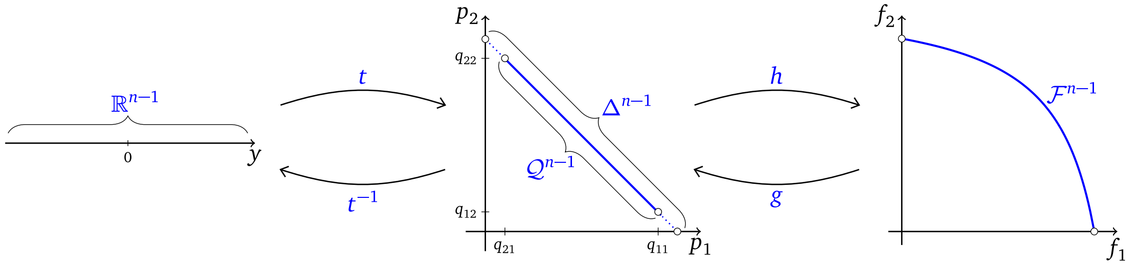

Figure 3 Sample space $\mathbb{R}^{n-1}$ (left), quasispecies space $\mathcal{Q}^{n-1} \subsetneq \Delta^{n-1}$ (center), and fitness space $\mathcal{F}^{n-1}$ (right). Sample space is mapped to the population distribution space by t with inverse t^{-1}. The quasispecies space $\mathcal{Q}^{n-1}$ is mapped to fitness space $\mathcal{F}^{n-1}$ by h with inverse g. Note the lower and upper bounds of the marginal relative frequencies in the two-haplotype model $q_{21} < p_1 < q_{11}$ and $q_{12} < p_2 < q_{22}$.

can be used for setting up the mutation matrix $\mathbf{Q}$ (File S1, section 1A).

Either way, the matrix $\mathbf{Q}$ is positive, symmetric, and stochastic. Furthermore, as $\mathbf{Q}$ is strictly diagonally dominant, it is also regular due to the Levy–Desplanques theorem (Horn and Johnson 1985). We fix $\mathbf{Q}$ for the rest of this article and focus on estimating the fitness landscape. In this setting, the quasispecies model is identifiable. By contrast, it has been shown that even with time series data, $\mathbf{Q}$ and $\mathbf{f}$ jointly are structurally nonidentifiable, because the quasispecies equation is overparameterized (Falugi and Giarré 2009).

While $\mathbf{Q}$ is mathematically irreducible, *i.e.*, every haplotype can mutate into every other haplotype with positive probability, technical precision limits imposed by machine precision can lead to situations where $\mathbf{Q}$ is not numerically irreducible any longer. Let $G_k = (\mathcal{H}, E_k)$ be the undirected graph with vertices $\mathcal{H}$ and edges $(i,j) \in E_k$ whenever $d(i,j) \leq k$. For a given k that depends on machine precision and on the mutation rate μ, $\mathbf{Q}$ will be numerically irreducible if the haplotype graph G_k is connected, *i.e.*, if a path exists between any pair of haplotypes. For HIV, we require $k \leq 1$ when using standard precision and $k \leq 2$ when using quadruple precision, for three reasons. First, biologically, any three mutations in one replication cycle occur with a probability of $\sim 10^{-15}$, which is orders of magnitude lower than the inverse of the number of virions. Second, the approximate transition rate of 10^{-15} is numerically bordering on machine $\varepsilon = 2.22 \times 10^{-16}$, leaving little overhead for precise calculations (see File S1, section 5A). Finally, most haplotypes will be generated by mutation from haplotypes closely related to them, such that in the asymptotic limit of an infinitely large population size, the great majority of the influx by mutation will still originate from very closely related haplotypes.

In practice, we construct $\mathcal{H}$ from observed data as follows (see Figure 2 for an illustration). For a given k (1 in the case of standard precision), we first construct G_k from the observed haplotypes and determine its connected components. If the number of connected components is >1, we proceed to augment $\mathcal{H}$ by including additional unobserved haplotypes to increase connectivity. We iterate over all pairs of connected components and determine the haplotype in each connected component closest to the haplotypes in the other connected component. We draw an edge between these two haplotypes with weight equal to their Hamming distance. We then build the minimum weight spanning tree between connected components. Finally, we replace all of the edges of the minimum weight spanning tree by linear chains of k mutational steps by inserting unobserved haplotypes. The stability of this procedure is detailed in File S1, section 5B.

Fitness landscape space: The quasispecies equation describes the dynamics of an evolving population of haplotypes, but in clinical practice, time series data are difficult and expensive to produce and thus scarce. Hence, we apply the model to cross-sectional data by analyzing the quasispecies in mutation–selection equilibrium (Equation 3). With this assumption, we cannot determine the timescale of approaching the equilibrium, which is reflected in the magnitude of the fitness landscape $\mathbf{f}$. We therefore constrain the average fitness to $\phi = 1$, removing 1 d.f. The constraint fitness space is denoted

$$\mathcal{F}^{n-1} = \{\mathbf{f} \in \mathbb{R}_+^n : \phi = \mathbf{p}^* \cdot \mathbf{f} = 1\}. \qquad (5)$$

We can now ask, for a given equilibrium distribution $\mathbf{p}$, what is the corresponding fitness landscape $\mathbf{f}$? This amounts to solving Equation 3 for $\mathbf{f}$. The solution defines the mapping $h : \mathcal{Q}^{n-1} \to \mathcal{F}^{n-1}$,

$$h(\mathbf{p}) = \mathbf{f} = \operatorname{diag}(\mathbf{p})^{-1}\mathbf{Q}^{-1}\mathbf{p}, \quad \mathbf{p} \in \mathcal{Q}^{n-1}, \qquad (6)$$

where $\mathcal{Q}^{n-1} \subsetneq \Delta^{n-1}$ is the quasispecies space defined above (File S1, section 2B). The mapping h is a bijection with inverse denoted by $g : \mathcal{F}^{n-1} \to \mathcal{Q}^{n-1}$ (Figure 3; File S1, sections 2 and 2A). This property is critical, as it allows for estimating fitness parameters in $\mathcal{F}^{n-1}$ from haplotype frequencies in $\mathcal{Q}^{n-1}$.

Inference of the fitness landscape

To estimate fitness in the equilibrium quasispecies model, we require a sample of the viral population. The data vector $\mathbf{X} \in \mathbb{N}^n$ records the count X_i of each haplotype i among $N = \sum_{i=1}^{n} X_i$ reads sampled in total. Classical methods such as limiting dilution assays with Sanger sequencing can be used to produce such data. Nowadays, NGS is far less laborious and produces data with ever increasing depth. As NGS data are generally noisy, *i.e.*, they include erroneously incorporated bases with respect to the true template, several methods have been developed to address this problem. Probabilistic and combinatorial approaches can be used for preprocessing raw NGS data to reduce errors and to infer the composition of the viral population $\mathbf{X}$ (Beerenwinkel *et al.* 2012).

If we assume that haplotypes have been inferred from raw sequencing data, then the read counts from one NGS experiment represent a sample from the multinomial distribution $\mathrm{Mult}(\mathbf{X}|\mathbf{p})$. Furthermore, if we assume $\mathbf{p} = g(\mathbf{f})$ as the quasispecies equilibrium distribution, this narrows the parameter space. Whereas usually Δ^{n-1} is the parameter space of the multinomial distribution, we have to work with quasispecies space $\mathcal{Q}^{n-1}$ as only elements from this set can represent true underlying quasispecies distributions. For certain data sets where $N^{-1}\mathbf{X} \notin \mathcal{Q}^{n-1}$, a maximum-likelihood estimator $\hat{\mathbf{p}}$ will not exist, because $\mathcal{Q}^{n-1}$ is an open set and is therefore not compact. Such situations can arise, for example, when $X_i = 0$ for some haplotype $i \in \mathcal{H}$. This property makes resampling techniques like bootstrapping intractable for realistic data sets due to the increasingly large number of bootstrap samples not having a maximum-likelihood estimator (MLE).

We address this statistical problem in a Bayesian fashion. This approach not only gives us the full posterior $p(\mathbf{f}|\mathbf{X})$ but also circumvents the aforementioned shortcomings of a likelihood-based approach. We employ a noninformative, maximum-entropy uniform prior on $\mathcal{F}^{n-1}$, by setting $p(\mathcal{F}^{n-1}) = \mathrm{const.}$ It remains to compute the posterior distribution

$$p(\mathbf{f}|\mathbf{X}) = \frac{P(\mathbf{X}\,|\,g(\mathbf{f}))p(\mathbf{f})}{P(\mathbf{X})}, \tag{7}$$

where $P(\mathbf{X}|g(\mathbf{f}))$ is the multinomial likelihood. As with most practical Bayesian inference problems, the posterior distribution cannot be derived in closed form. This is due to the eigenvalue constraint in Equation 5, which is a nonlinear constraint and thus yields the nonlinear bounded space $\mathcal{F}^{n-1}$.

MCMC sampler: We have developed a Metropolis–Hastings MCMC sampler that can draw samples from this posterior distribution. This task is daunting, because working directly on the fitness space $\mathcal{F}^{n-1}$ would require knowledge of the neighborhood structure of this nonlinear space. Furthermore, working on $\mathcal{Q}^{n-1}$ is also difficult as, in general, its boundary is not analytically known. Finally, the most common distributions on the simplex, such as the Dirichlet distribution, have strong conditional independence properties that do not allow for flexible covariance structures (Aitchison and Shen 1980).

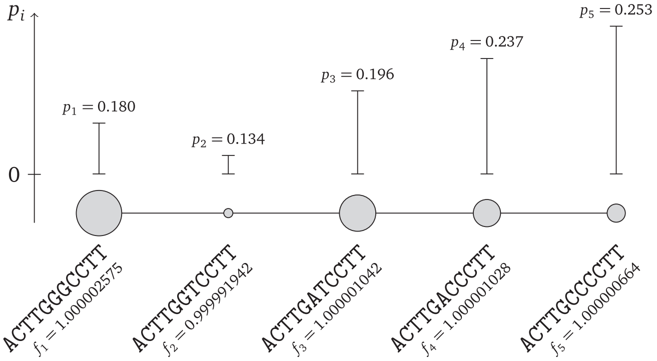

Figure 4 Five-haplotype simulation model. The haplotype graph G_1, where haplotypes can mutate into all other haplotypes taking only steps of one mutation at a time, is a chain. The true fitness ranks are indicated by the size of the nodes and the equilibrium relative abundance by the height of the vertical line above each haplotype node.

We use $\mathbb{R}^{n-1}$ as our sampling space and map samples to the fitness manifold $\mathcal{F}^{n-1}$ via $\mathcal{Q}^{n-1}$, using the composed mapping $h \circ t$, where $t : \mathbb{R}^{n-1} \to \Delta^{n-1}$ is the logistic transformation defined by $t_i(\mathbf{y}) = \exp(y_i)(1 + \sum_{j=1}^{n-1} y_j)^{-1}$, for $i = 1, \ldots, n-1$, and $t_n(\mathbf{y}) = (1 + \sum_{j=1}^{n-1} y_j)^{-1}$ (Figure 3; File S1, section 4). The approach of mapping samples from a simpler space to a manifold has been demonstrated by Diaconis *et al.* (2013). Working in Euclidean space is much easier and we have more distributions at our disposal for constructing proposal distributions. For the functional form of the posterior we have, up to a normalization constant,

$$\log p\,(\mathbf{y}|\mathbf{X}) = \log d(\mathbf{y}) + \sum_{i=1}^{n} X_i \log p_i + \mathrm{const.} \quad \mathbf{y} \in \mathbb{R}^{n-1}, \tag{8}$$

where $d(\mathbf{y}) = |\det(\mathbf{J}[h \circ t])|$ and $\mathbf{J}$ denotes the Jacobian (see File S1, section 3).

One widely characterized class of fitness landscapes is Stuart Kauffman's LK fitness landscapes (Kauffman and Weinberger 1989), where L denotes the number of genomic loci and K denotes the number of interacting loci of each locus. Campos *et al.* (2002) have shown that the only LK fitness landscapes that lack any correlation between closely related haplotypes are the house of cards fitness landscapes, *i.e.*, fitness landscapes where all genes interact with all other genes concurrently. Thus, it is natural to assume at least some degree of correlation inherent in real fitness landscapes.

Capturing any potential correlation present in fitness landscapes is key to an efficient Bayesian estimation of the posterior. To this end, we employ a differential evolution MCMC as a sampling algorithm. This sampler is of the globally adaptive type and can estimate the covariance structure efficiently. It has the advantage of requiring no *a priori* specification of the posterior's covariance structure, which is generally not known. Differential evolution MCMC (Ter Braak 2006) is very similar in nature to parallel tempering (Earl and Deem 2005) for estimating equilibrium distributions of energy states. The basic idea is that the difference of two chains of the current population optimally captures the correlation structure. An advantage of our implementation over other globally adaptive MCMC schemes is the ease of parallelizing this otherwise iterative procedure. Furthermore, due to crossover of chains, the inference scheme involves minimal overhead computation.

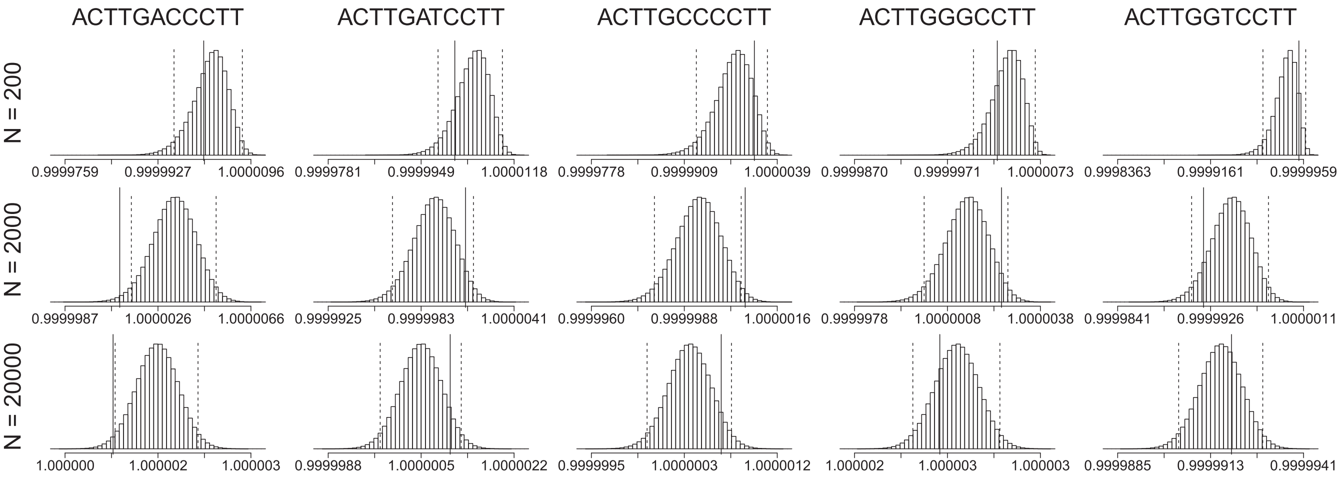

Figure 5 Posterior fitness distributions of each of the five haplotypes in the simulations. Each histogram displays the marginal posterior distributions of a simulation run with a particular coverage *N* indicated to the left of each row. The true parameter f_i of each haplotype from Figure 4 is indicated by a solid vertical line. The 95% highest posterior density intervals are demarcated with dashed vertical lines. The true parameter is not included in the 95% highest posterior density intervals in some cases, as is to be expected from a finite sample.

In practice, for an n-dimensional problem, we run $\sim 2n$ independent chains. At every generation, we cycle through all chains and update the current chain by generating a new proposal. For this, we calculate the proposal for chain i with $\mathbf{y}_p = \mathbf{y}_i + \gamma(\mathbf{y}_{R_1} - \mathbf{y}_{R_2}) + \mathbf{z}$, where $R_1 \neq R_2$ denote two randomly chosen indexes of vectors of chains other than i in each trial, and $\mathbf{z} \sim \mathcal{N}(0, \sigma\mathbb{I}_n)$ is responsible for detailed balance to hold. This proposal sample is retained only if all its components are positive by inspection of Equation 6. If the sample has not been rejected in the previous step, then we calculate the log posterior (Equation 8) and determine by the usual MCMC acceptance probability $\min\{1, p(\mathbf{y}_p|\mathbf{X})/p(\mathbf{y}_i|\mathbf{X})\}$ whether to accept this sample as a draw from the posterior. We start the parallel chains at the MLE if it exists; otherwise we start near the boundary where the MLE would be if the parameter space was closed.

Implementation

We implemented our MCMC inference scheme, called *QuasiFit*, in C++. For the linear algebra we employed the Eigen suite (Guennebaud and Jacob 2010), which is a flexible framework for calculating Lower Upper (LU) decompositions that are crucial for matrix inversion and for calculating the determinant of a matrix.

Convergence of all MCMC runs was assessed with the coda package in R (Plummer *et al.* 2006). For the clinical data sets, we plotted the Gelman and Rubin scale reduction factor *vs.* trial number and the autocorrelation, and we tested for equality of distribution in the purported stationary distribution samples (File S1, section 6).

The major computational bottleneck in our inference scheme is the calculation of the determinant $\log d(\mathbf{y})$ in each step. The computational complexity is therefore $\mathcal{O}(n^3)$ for every trial due to the LU decomposition. In total this makes for a complexity of $\mathcal{O}(N_{\text{trials}} \cdot n^3)$. We conducted simulations to investigate the required central processing unit (CPU) time per MCMC trial (File S1, section 9). In practice, the asymptotic regime is reached for $n > 64$. If all considered haplotypes have at least one observation, *i.e.*, $X_i > 0$ for all $i \in \mathcal{H}$, then we have the best-case decline in efficiency of $\mathcal{O}(n^{-1})$ in Metropolis–Hastings schemes (Hanson and Cunningham 1998).

The major factor for convergence, well-mixing and statistical efficiency is determined by the number of haplotypes with no observations ($X_i = 0$). In practice, this condition is identical to asking for the existence of a MLE. In the case where some $X_i = 0$, an MLE does not exist. This in turn will lead to a situation where the posterior on $\mathcal{Q}^{n-1}$ will be located on the boundary. The more haplotypes are considered with $X_i = 0$, the longer it will take for the initial burn-in phase to approach the boundary and the less efficient the overall procedure becomes. Furthermore, once in the stationary distribution at the boundary, with increasing number of haplotypes without observations an increasing number of rejections in the Metropolis–Hastings schemes result not from the general curse of dimensionality of MCMC schemes but from proposal samples that are not elements of $\mathcal{Q}^{n-1}$. In practice, data sets of up to 300 haplotypes can currently be analyzed on a 48-core system.

Results

Simulation studies

As there are no known *in vivo* fitness values of viruses from the same host, we resorted to simulations to assess the goodness of our fitness landscape estimates.

Five-haplotype simulation model: We first devised a small five-haplotype example to illustrate the intricate relationship between haplotype fitness and frequency. Figure 4 shows the haplotype network G_1 and parameters used for

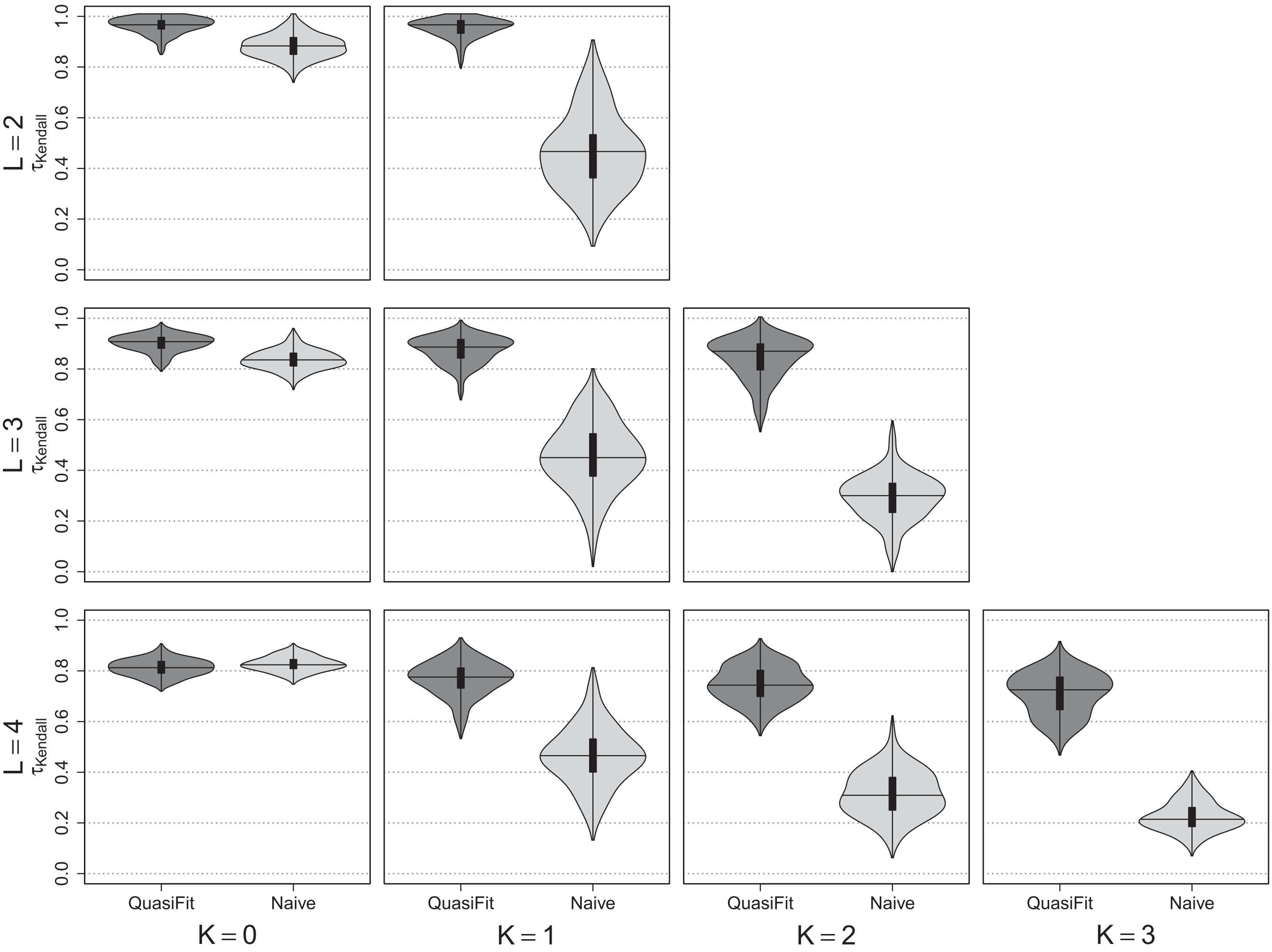

Figure 6 Rank correlation coefficient τ_{Kendall} for different L and K. The rows depict results for increasingly high-dimensional DNA spaces, where $n = 4^L$ denotes the number of haplotypes. The columns depict the density estimators for the rank correlation coefficients between an estimator and the true fitness landscape with increasing K. Densities with dark shading represent τ_{Kendall} for the *QuasiFit*-based estimator and densities with light shading represent τ_{Kendall} for the count-based estimator.

the simulation. We drew multinomial samples with different read coverages $N \in \{200, 2000, 20000\}$ and applied our MCMC approach to obtain the marginal posterior fitness distributions shown in Figure 5. As coverage N increases, so does the confidence in the estimated fitness values, which manifests itself in smaller credibility intervals. The most fit haplotype is only the second least frequent and the haplotype with the highest frequency is the second least fit. This weak correlation between the ranks of relative abundances and fitness is an important consequence of mutational coupling. It highlights the potential pitfalls of simply relying on abundance as a measure of fitness.

LK fitness landscape simulations: To validate more realistic fitness landscapes, where n is at least on the order of the expected number of viral haplotypes in patient samples, as presented below, we employed the LK model for simulating fitness landscapes. Stuart Kauffman's LK model (originally called the NK model) is a widely used scheme for generating random fitness landscapes of tunable ruggedness (Kauffman and Weinberger 1989; Szendro *et al.* 2013).

The LK model is defined such that L determines the total number of loci and $K < L$ determines the number of other loci affecting the fitness of any one allele at some locus. Let $\mathbf{a} = (a_1, \ldots, a_L) \in \mathcal{A}^L$ be a DNA sequence of length L and $\mathbf{e}_i = \{i, e_{i,1}, e_{i,2}, \ldots, e_{i,K}\}$ be the interaction structure of locus i, *i.e.*, the K other loci affecting locus i. The fitness landscape $\mathbf{f} : \mathcal{A}^L \rightarrow \mathbb{R}_+$ is then

$$\mathbf{f}(\mathbf{a}) = \sum_{i=1}^{L} b_i\left(\left(a_j\right)_{j\in\mathbf{e}_i}\right), \qquad (9)$$

where $b_i(\cdot)$ denotes the fitness contribution of allele a_i. To generate a random fitness landscape according to the LK model with given L and K, we proceed by first generating random interaction sites. For every locus i, we randomly select K elements from $\{1, \ldots, i - 1, i + 1, \ldots, N\}$ (without replacement) as the loci $\mathbf{e}_i$ on which the fitness at locus i

depends. Second, we generate the mapping $b_i((a_j)_{j\in\mathbf{e}_i})$ by randomly sampling values from $\log\mathcal{N}(\eta,\nu)$, where the parameters η and ν determine the mean and, respectively, the standard deviation of the log-normal distribution. We set $\eta = 10^{-2}$ and $\nu = 2 \times 10^{-6}$ to produce fitness values with small differences between each other such that the quasispecies shows significant diversity in equilibrium.

For every pair of $L \in \{2, 3, 4\}$ and $K \in \{0, \ldots, L-1\}$, we generated random fitness landscapes and calculated the stationary distribution, sampled one multinomial sample with simulated read coverage of $N = 100{,}000$, and repeated this procedure until we had 100 samples possessing a fitness MLE. It should be emphasized here that our inference scheme does not require an MLE (File S1, section 5B). Requiring all samples to have an MLE was solely done to facilitate convergence for high-dimensional haplotype sets with $L = 4$. Finally, we ran *QuasiFit* on all multinomial samples and took the mean of the posterior as an estimator of the underlying fitness landscape.

To compare our model-based predictions to those of merely using the ranks of the estimated frequencies as a proxy for the ranks of the fitness landscape, we used Kendall's τ as a measure of agreement in the ranks of different methods of estimating fitness landscapes (Figure 6). The case $K = 0$ represents fitness landscapes possessing only main/additive effects; *i.e.*, there is no epistasis and as such we can envision a Mount Fuji-like fitness landscape, where mutations will cause the population to ultimately climb to the maximum fitness, as there is only one local optimum that is also the global optimum. The ranks of the fitness landscape and its equilibrium population distribution closely agree in this case (Figure 6). For $K > 0$, our fitness landscape estimates recover the ranks of the true fitness landscape significantly better than the naive count-based estimator as our model accounts for mutational neighborhood structure (Figure 6).

Next, we assessed the ability of our model to estimate the rank fitness landscape as a function of the magnitude of epistatic relative to additive effects. We decomposed the LK fitness landscape model into its parametric interaction terms and simulated random fitness landscapes with prescribed epistatic strengths, using interactions of order at most $K + 1$. We found that our model becomes superior to the naive ranking method as soon as epistatic effects are on the order of 10% of the additive effects, which is well within a biologically plausible range of epistatic effects in HIV (File S1, section 5E).

Robustness of mutation parameter μ: To assess the robustness of our inference scheme with respect to the presumed mutation rate, we performed a large-scale analysis over a range of mutation rates, $10^{-6} \leq \mu \leq 10^{-3}$. To be realistic in view of the clinical data and rudimentary current knowledge of fitness landscapes, we simulated one fitness landscape on a DNA space of $L = 3$ (64 haplotypes) and with some epistasis $K = 1$, using the procedure outlined in the previous section. The choice of the inclusion of first-order epistatic interactions is motivated by the analysis in Hinkley *et al.* (2011), where such pairwise interactions (equivalent to $K = 1$) have been found to be an important feature of HIV-1 fitness landscapes.

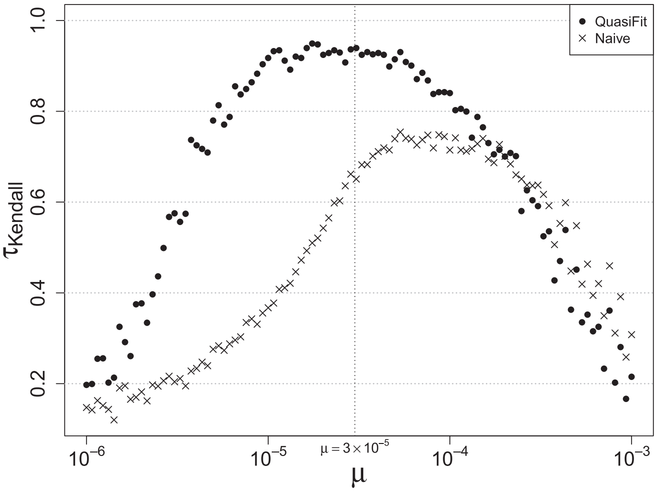

Figure 7 Rank correlation coefficient τ_{Kendall} between the true fitness landscape and one of the two estimators plotted against different actual simulated mutation rates. The dotted vertical line marks the mutation rate that we assume in our model.

We generated the fitness landscape with parameters $b_i(\cdot) \sim \log\mathcal{N}(\eta,\nu)$, where $\eta = 10^{-2}$ and $\nu = 5 \times 10^{-7}$. The latter parameter is one-fourth of the corresponding ν in the previous LK simulations such that the average selective advantages produced in the fitness landscape are not much larger than the mutation rate at the lower bound of 10^{-6}; otherwise the coupling between haplotypes is too weak and no diversity will be present at equilibrium. We have iterated over 100 log-uniformly spaced μ-values in the interval $[10^{-6}, 10^{-3}]$. For each value of μ, we calculated the equilibrium distribution given our fitness landscape and the mutation matrix (Equation 4). For each equilibrium distribution, we simulated a read coverage of $N = 100{,}000$ by drawing from a multinomial distribution with $\mathbf{p} = g(\mathbf{f})$ and then applying our sampler with fixed $\mu = 3 \times 10^{-5}$. We sampled a total of $N_{\text{trials}} = 43.2 \times 10^6$ with 144 chains and a thinning interval of 100, giving us 432,000 samples after each run for every μ. We calculated the mean marginal fitnesses of the last 100,000 samples and determined the rank correlation coefficient τ_{Kendall} with respect to the initially fixed true fitness landscape for the *QuasiFit*-based estimator and the naive count-based estimator. The actual mutation rates *vs.* τ_{Kendall} for the back-inference and the naive estimator are shown in Figure 7.

We found our model to be very robust within half an order of magnitude below and above our presumed mutation rate of $\mu = 3 \times 10^{-5}$. Our estimates also reproduce the ranks of the true fitness landscape better than the naive estimator over a wide range of μ-values. In general, reproducing a perfect agreement (*i.e.*, $\tau = 1$) between ranks of the true and reinferred fitness landscapes is not possible,

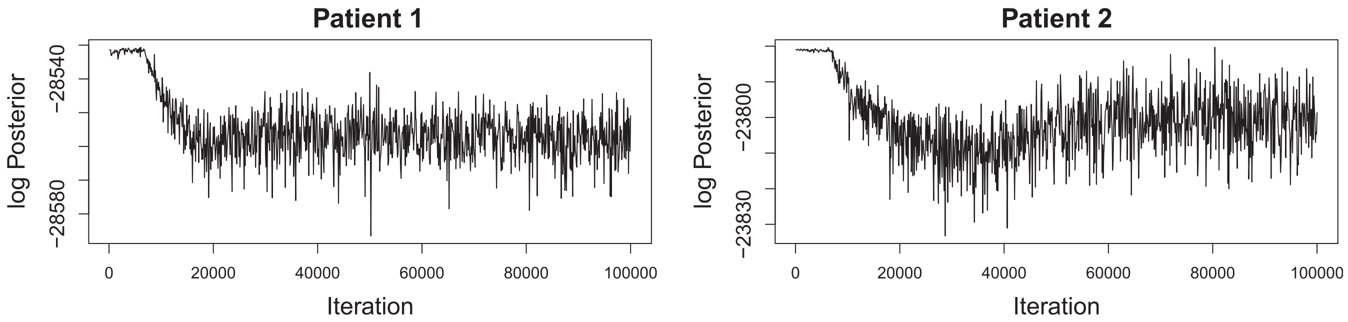

Figure 8 Log-posterior of the first 100,000 MCMC trials. The *x*-axis denotes the trial number and the *y*-axis denotes the logarithm of the posterior distribution, Equation 8.

due to the introduced sampling variance of the finite multinomial draw for each μ.

We also conducted simulations to assess the robustness toward violations of the transition/transversion rate κ (File S1, section 5D). These simulations suggest that our inference scheme still predicts better ranks of the fitness landscape when κ shows large deviations from 1.

Sensitivity analysis: In addition, we performed a sensitivity analysis that shows our method to be significantly better at recovering the ranks of the underlying fitness landscape up to 500 time units away from equilibrium (File S1, section 5C).

Fitness landscapes of clinical p7 quasispecies

To apply our model to clinical data, we selected two patients from the Swiss HIV Cohort Study (Schoeni-Affolter *et al.* 2010). We analyzed parts of the spacer peptide 1 (p2) and the nucleocapsid protein (p7) comprising a total of 207 bases or 69 amino acids. We aligned the 2 × 250-bp Illumina MiSeq reads with *bwa* (Li and Durbin 2010) to the reference sequence HXB2 and focused on the p2–p7 reading frame. The compositions of the viral populations had been inferred with the probabilistic viral haplotype reconstruction tool *QuasiRecomb* (Töpfer *et al.* 2013). In both cases, *QuasiRecomb* was employed only for error correction, as the raw reads contain too many errors, but read assembly was not necessary

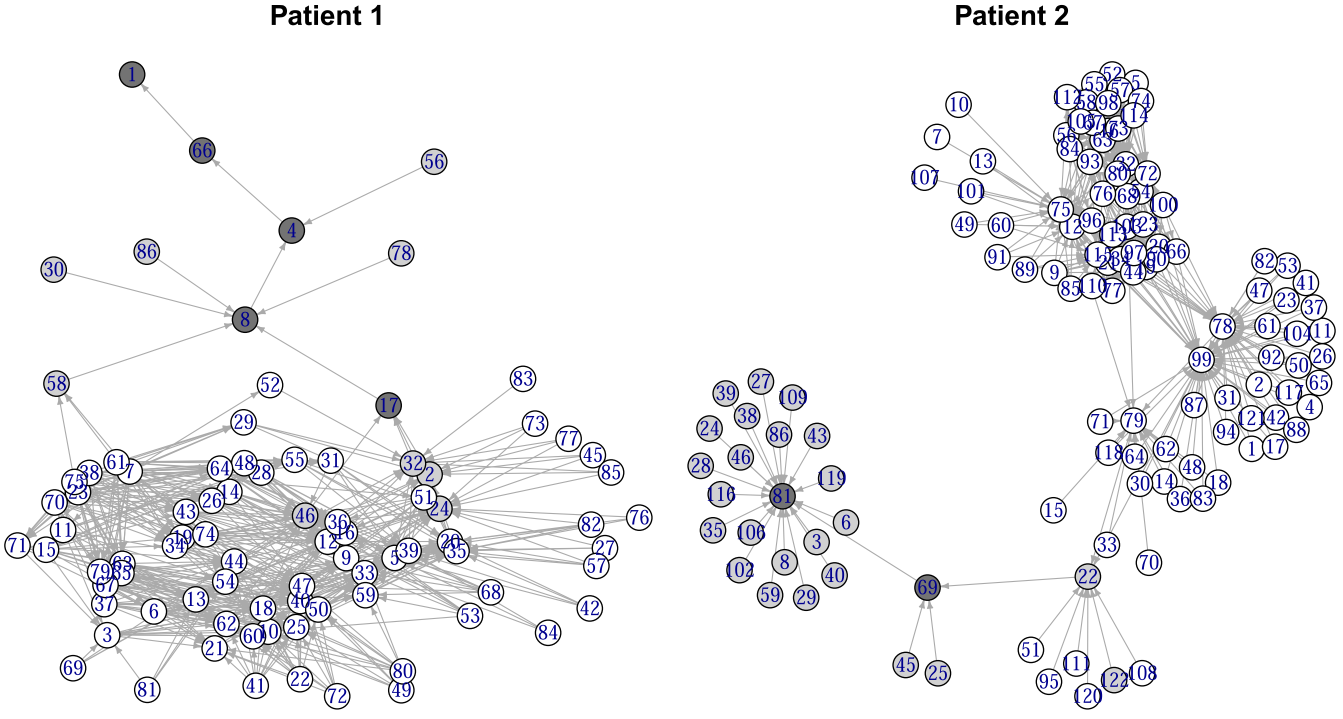

Figure 9 Rank fitness landscapes of the haplotypes in each of two patients. A directed edge $i \rightarrow j$ exists between haplotypes *i* and *j* if the posterior fitness difference $f_j - f_i$ can credibly be inferred to be >0, *i.e.*, if, given the model, there is evidence for haplotype *j* being fitter than haplotype *i*. Both graphs possess the transitive property; *i.e.*, if *j* is fitter than *i* (indicated by an edge $i \rightarrow j$) and *k* is fitter than *j*, then *k* is also fitter than *i* and a directed path exists from *i* to *k*. Dark gray vertices possess credibly larger than average, light gray vertices possess average, and colorless vertices possess lower than average fitness $\phi = 1$.

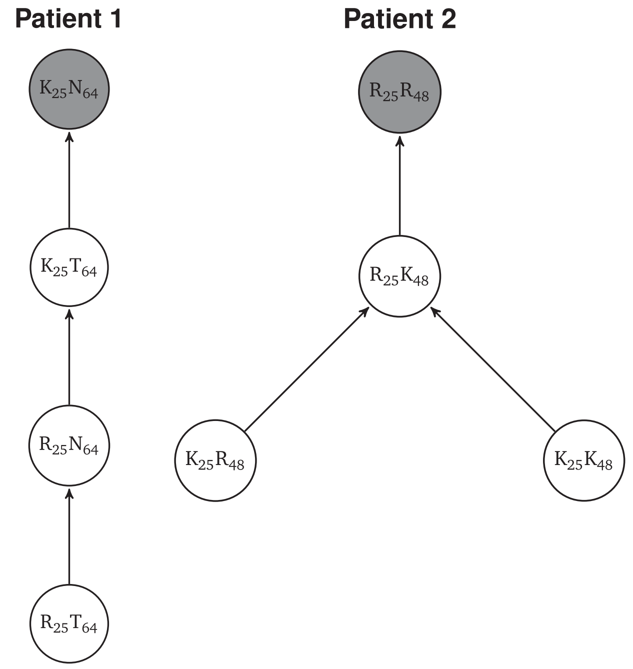

Figure 10 Rank fitness landscapes of four peptides in each of two patients. Each peptide is denoted by amino acids subscripted with their loci. A directed edge $i \rightarrow j$ exists between peptides i and j if the posterior fitness difference $F_j - F_i$ can credibly be inferred to be >0, *i.e.*, if, given the model, there is evidence for peptide j being fitter than peptide i. Vertices with dark shading possess larger than average and open vertices possess lower than average fitness $\phi = 1$.

on these short segments. After inference, the observed quasispecies distribution for patient 1 and patient 2 consisted of $n_1 = 86$ and, respectively, $n_2 = 123$ haplotypes. Both data sets possess a connected haplotype graph G_1, such that no further unobserved haplotypes needed to be included.

We determined the fitness landscape of the quasispecies by running *QuasiFit* on the estimated composition of the viral population. Default parameters were used. In total, 288 chains were run in parallel, with a total of 500,000 trials per chain for both patient samples. The burn-in phase for both MCMC processes was ~30,000 and 60,000 (Figure 8) and every chain was thinned by retaining every 1150th sample (File S1, Figure S6 and section 6). In Figure 8, the observed drop from the initial value is due to the general curse of dimensionality of sampling in high dimensions when starting at the single highest point of the posterior.

We determined neutral haplotype networks by 95% highest posterior density (HPD) intervals of marginal fitness differences. A 95% HPD region is the smallest region that has 95% probability mass. We determine these for the marginals of the posterior by minimizing over all 95% credibility intervals. Neutrality between two haplotypes is called when 0 is an element of the 95% HPD of fitness differences between two such haplotypes. On the other hand, a ranking of haplotypes by fitness can be established when there is a credible difference in fitnesses, *i.e.*, when 0 is not an element of the pairwise fitness difference (Figure 9; haplotype sequences in File S1, section 7). Visualizing fitness landscapes by drawing a directed edge between haplotypes of differing fitness is a popular and intuitive way of visualizing these high-dimensional mathematical objects (Crona *et al.* 2013). Both graphs are transitively reduced; *i.e.*, all directed paths between haplotypes represent credible fitness differences. In patient 1, the fitness landscape is dominated by a few highly fit haplotypes and a large fraction of unfit haplotypes. In contrast, patient 2 shows a fitness landscape where the two highly fit haplotypes are surrounded by a cloud of haplotypes of intermediate, average fitness. It also shows a stronger star-like topology in comparison to that of patient 1.

To summarize the haplotype networks of size $n_1 = 86$ and $n_2 = 123$, we translated their DNA sequences into peptides and calculated the joint posterior fitness distribution of each peptide i as $F_i = (\sum_j f_j p_j)/(\sum_j p_j)$, where the sums run over all DNA sequences j that code for peptide i. We discarded all loci with conserved residues and denote peptides with their alleles subscripted by their loci. Figure 10 illustrates the fitness landscapes for the four peptides in patients 1 and 2. This analysis shows again the importance of working with fitness values instead of ranking frequencies. In patient 1, while $K_{25}T_{64}$ and $R_{25}N_{64}$ show no credible differences when comparing their posterior frequencies, *i.e.*, zero is an element of the 95% HPD region, they do show a credible difference in their fitness values, *i.e.*, zero is not an element of the 95% HPD region (Figure 11).

In addition to pairwise fitness differences, we also used our method for detecting epistasis, *i.e.*, nonadditive effects of multiple alleles on fitness. We denote with 0 the wild-type or major allele and with 1 the mutant or minor allele. In patient 1, K_{25} and N_{64} are the major alleles and in patient 2, R_{25} and R_{48} are the major alleles. To determine whether a nonlinear interaction exists between alleles at two loci, we considered the random variable $\delta = F_{11} + F_{00} - F_{10} - F_{01}$ and tested whether δ is credibly different from 0. In the clinical data, we found a credible, nonzero epistatic effect between loci 25 and 64 in patient 1, but not between loci 25 and 48 in patient 2 (Figure 12).

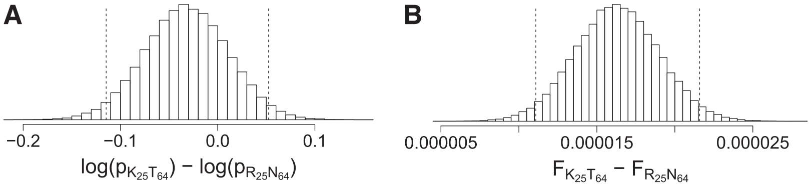

Figure 11 (A and B) Histograms of the difference of log-marginal frequencies (A) and of marginal fitness values (B). Differences in fitness values need not necessarily correspond to differences in marginal frequencies. The differences here are based on peptides from patient 1. The 95% highest posterior density intervals are demarcated with dashed vertical lines.

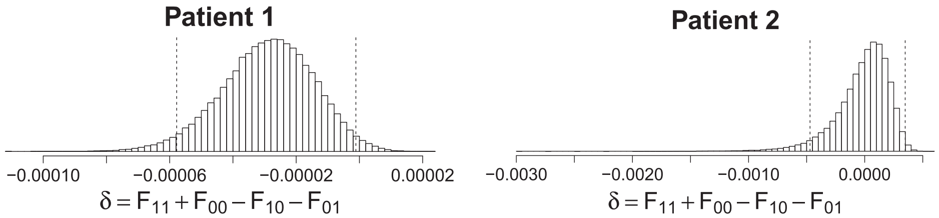

Figure 12 Posterior distributions of the epistasis term δ in both patients. A credible interaction term was detected in patient 1, but not in patient 2. The 95% highest posterior density intervals are demarcated with dashed vertical lines.

Finally, we also analyzed codon usage effects. We found that the marginal fitness differences between major and minor codons at synonymous amino acid residues were all credible, mainly due to the large difference in their frequencies (File S1, section 8).

Discussion

We have developed a computational framework for inferring fitness landscapes from NGS samples of HIV-1 patient-derived viruses, using quasispecies theory. Our inference scheme represents a novel approach to derive a measure of fitness from cross-sectional data. We obtained unimodal posterior distributions because of the existence of a single global stable mutation–selection equilibrium in the quasispecies model. This makes the inference scheme particularly efficient, as multimodality and suboptimal exploration of fitness space are not a problem. Our inference scheme is strongly parallelizable. For instance, for 300 haplotypes, we can run 624 chains on a 48-core server. Then each core needs to update only 13 chains by itself. This efficient partitioning scheme for sampling cannot be attained with ordinary globally adaptive MCMC schemes, where much CPU power cannot be utilized due to inherent lack of concurrency.

Every evolutionary model includes assumptions to make analysis possible. While established in the field of virology with many ubiquitous applications, the quasispecies framework nonetheless has limitations. One central weakness of the quasispecies model is at the same time its greatest strength, namely the assumption of a constant fitness landscape. Due to the inherent feedback loop of the host immune system, the quasispecies model will likely fail in genomic regions that can experience strong immunological pressure such as *env*, where fitness is more likely to be time dependent. While adaptations exist of the quasispecies model to time-varying, frequency-dependent fitness landscapes, such approaches necessarily include more involved mathematical machinery. The replicator–mutator equation is one such extension. In light of the minimum n^2-dimensional parameter space and the highly complicated nonlinear trajectories of the replicator equation coupled with mutation (Pais and Leonard 2011), such a model would require prohibitively large amounts of time series, rather than cross-sectional, data to be useful for inference.

The centerpiece of quasispecies theory is the quasispecies, *i.e.*, the population in mutation–selection equilibrium. Whether in reality such an equilibrium can ever exist remains an open question in the field of virology. One early study by Domingo *et al.* (1978) supported such a dynamic equilibrium for a multiply passaged Qβ bacteriophage. Ramratnam *et al.* (1999) highlight the existence of a dynamic equilibrium of production and clearance of HIV particles *in vivo*. Quasispecies theory likely fails to account for the acute phase of HIV infection, which is characterized by very strong initial immune responses that will show strong dynamics and where coupling between haplotypes is less of a driving force than immune escape.

Due to the high rates of mutation in RNA viruses and their large population sizes, quasispecies theory makes the implicit assumption that the expected number of produced mutants per replication cycle $N_e\mu$ is large and hence can be modeled quasi-deterministically (Rodrigo 1999). It should be noted that the precise value of N_e in viral populations is an open question. In particular, if $N_e < \mu^{-1}$, the evolutionary process is dominated by stochastic effects, such that genetic drift trumps deterministic forces like selection. In this stochastic regime, the informative value of one sample of a viral population diminishes rapidly with decreasing N_e. Given these large random fluctuations in the stochastic regime, inferring selection requires multiple replicates of time series data, to disentangle deterministic effects from random fluctuations. At least one study of linkage disequilibrium in HIV-1 suggests N_e to be $>\mu^{-1}$ (Rouzine and Coffin 1999). The N_e limitation does not just affect the quasispecies model, but all deterministic models of virus evolution, such as those based on ordinary differential equations.

In addition to the presumed quasispecies, we do not take recombination into account. While extensions of the quasispecies model exist that account for this phenomenon (Boerlijst *et al.* 1996; Jacobi and Nordahl 2006), they are exceedingly complicated by nonlinear dynamics arising from bimolecular production reactions. This generally leads to bistability, such that a unique global quasispecies is not guaranteed anymore. Here, we analyzed genomic regions for which we assume recombination within the region to be negligible but recombination outside of the region may be somewhat larger, such that genetic variation that exists outside of the region of interest does not confound the analysis. This is reasonable, as the genomic ranges analyzed here are <200 bp. In general though, the recombination rate will affect epistasis and above a certain threshold, allelic selection takes over, where the selective advantage of alleles depends only on their own intrinsic fitness contribution and not on combinations with other alleles (Neher and Shraiman 2009).

Every computational approach to inference of high-dimensional data includes certain assumptions and approximations necessary for making practical analysis possible. Our approach to analyzing NGS data in the form of an MCMC sampler is no different in this regard. Our inference scheme does not yet capture overdispersion, the effect of the data having more variance than postulated by the statistical model. It is well known that NGS data involve a number of experimental steps for sample preparation. Sequencing has practical limitations and is error prone. All these steps will eventually yield a population sample that is overdispersed with respect to the multinomial distribution we employ. However, without replicates, technical overdispersion cannot be estimated jointly with the fitness landscape. An alternative would be to specify experimental overdispersion upfront as additional model parameters.

Inferring *in vivo* fitness landscapes for comparative analysis becomes possible with our inference approach. It could prove to be fruitful with regard to fitness landscapes of, for example, genomic loci that experience negligible immunological pressure and are not subject to drug pressure. One example is HIV-1's *gag* gene, parts of which we have analyzed here. We can determine reliably the number of distinct fitness classes. In the analyzed patient data, we deduced a total order of peptides with increasing fitness in patient 1 and a partial order in patient 2. The reason for not being able to infer a total order in the case of patient 2 lies in the neutral network formed by the peptides $K_{25}R_{48}$ and $K_{25}K_{48}$. Additionally, we can also analyze properties of fitness landscapes, such as epistatic interactions.

In clinical settings, the vast majority of data will be cross-sectional and not time series. To go beyond the current standard of practice of equating fitness ranks to frequency ranks, any fitness inference method based on cross-sectional data will need to make strong assumptions on the population dynamics $\dot{\mathbf{p}}(t)$. By accounting for mutational neighborhood structure, an important factor of intrahost viral evolution (Burch and Chao 2000), our model performs significantly better at inferring the rank fitness landscape than equating fitness ranks to frequency ranks. This is important in light of the observation that fitness need not necessarily show a strong connection to relative haplotype abundance (De la Torre and Holland 1990).

In summary, we have devised a mathematical framework based on the quasispecies model and an efficient sampling scheme for estimating *in vivo* viral fitness from intrahost NGS data. It will help in analyzing viral populations and understanding their evolutionary dynamics and eventually their clinical consequences.

Acknowledgments

This work was supported by ETH research grant ETH-33 13-1 (to N.B.), by the University of Zurich's Clinical Research Priority Program "Viral infectious diseases: Zurich Primary HIV Infection Study" (H.F.G.), and by the Swiss National Science Foundation under grant CR32I2_146331 (to N.B., K.J.M., and H.F.G.).

Literature Cited

Acevedo, A., L. Brodsky, and R. Andino, 2014 Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505(7485): 686–690.

Aitchison, J., and S. M. Shen, 1980 Logistic-normal distributions: some properties and uses. Biometrika 67(2): 261–272.

Anderson, J. P., R. Daifuku, and L. A. Loeb, 2004 Viral error catastrophe by mutagenic nucleosides. Annu. Rev. Microbiol. 58: 183–205.

Beerenwinkel, N., L. Pachter, and B. Sturmfels, 2007a Epistasis and shapes of fitness landscapes. Stat. Sin. 17: 1317–1342.

Beerenwinkel, N., L. Pachter, B. Sturmfels, S. F. Elena, and R. E. Lenski, 2007b Analysis of epistatic interactions and fitness landscapes using a new geometric approach. BMC Evol. Biol. 7(1): 60.

Beerenwinkel, N., H. F. Günthard, V. Roth, and K. J. Metzner, 2012 Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3: 329.

Beerenwinkel, N., H. Montazeri, H. Schuhmacher, P. Knupfer, V. von Wyl *et al.*, 2013 The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients. PLoS Comput. Biol. 9(8): e1003203.

Boerlijst, M. C., S. Bonhoeffer, and M. A. Nowak, 1996 Viral quasispecies and recombination. Proc. R. Soc. Lond. B Biol. Sci. 263 (1376): 1577–1584.

Burch, C. L., and L. Chao, 2000 Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406(6796): 625–628.

Campos, P. R., C. Adami, and C. O. Wilke, 2002 Optimal adaptive performance and delocalization in NK fitness landscapes. Physica A 304(3): 495–506.

Clavel, F., and A. J. Hance, 2004 HIV drug resistance. N. Engl. J. Med. 350(10): 1023–1035.

Crona, K., D. Greene, and M. Barlow, 2013 The peaks and geometry of fitness landscapes. J. Theor. Biol. 317: 1–10.

De la Torre, J., and J. Holland, 1990 RNA virus quasispecies populations can suppress vastly superior mutant progeny. J. Virol. 64(12): 6278–6281.

Deforche, K., R. Camacho, K. Van 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(1): 34–41.

Diaconis, P., S. Holmes, M. Shahshahani, 2013 Sampling from a manifold, edited by T. Seppäläinen, K. Burdzy, S. Fienberg, P. Hall, R. Li, P. Green, A. DasGupta, and T.N. Sriram, pp. 102–125 in *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*. Institute of Mathematical Statistics. Beachwood, OH.

Domingo, E., D. Sabo, T. Taniguchi, and C. Weissmann, 1978 Nucleotide sequence heterogeneity of an RNA phage population. Cell 13(4): 735–744.

Domingo, E., J. Sheldon, and C. Perales, 2012 Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. 76(2): 159–216.

Earl, D. J., and M. W. Deem, 2005 Parallel tempering: theory, applications, and new perspectives. Phys. Chem. Chem. Phys. 7(23): 3910–3916.

Eigen, M., and P. Schuster, 1977 A principle of natural self-organization. Naturwissenschaften 64(11): 541–565.

Eigen, M., J. McCaskill, and P. Schuster, 1988 Molecular quasispecies. J. Phys. Chem. 92(24): 6881–6891.

Falugi, P., and L. Giarré, 2009 Identification and validation of quasispecies models for biological systems. Syst. Control Lett. 58(7): 529–539.

Ferguson, A. L., J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker *et al.*, 2013 Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity 38(3): 606–617.

Gavrilets, S., 2004 *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, NJ.

Guennebaud, G., and B. Jacob, 2010 Eigen v3. Available at: http://eigen.tuxfamily.org.

Hanson, K. M., and G. S. Cunningham, 1998 Posterior sampling with improved efficiency, edited by K. M. Hanson, pp. 371–382 in *Medical Imaging'98*. International Society for Optics and Photonics. Bellingham. WA.

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. Nat. Genet. 43(5): 487–489.

Horn, R. A., and C. R. Johnson, 1985 *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

Jacobi, M. N., and M. Nordahl, 2006 Quasispecies and recombination. Theor. Popul. Biol. 70(4): 479–485.

Kauffman, S. A., and E. D. Weinberger, 1989 The NK model of rugged fitness landscapes and its application to maturation of the immune response. J. Theor. Biol. 141(2): 211–245.

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(3): e1002551.

Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26(5): 589–595.

Ma, J., C. Dykes, T. Wu, Y. Huang, L. Demeter *et al.*, 2010 vFitness: a web-based computing tool for improving estimation of in vitro HIV-1 fitness experiments. BMC Bioinformatics 11(1): 261.

Metzner, K. J., S. G. Giulieri, S. A. Knoepfel, P. Rauch, P. Burgisser *et al.*, 2009 Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. Clin. Infect. Dis. 48(2): 239–247.

Musso, F., 2012 On the relation between the Eigen model and the asexual Wright–Fisher model. Bull. Math. Biol. 74(1): 103–115.

Neher, R. A., and B. I. Shraiman, 2009 Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc. Natl. Acad. Sci. USA 106(16): 6866–6871.

Niedringhaus, T., D. Milanova, M. Kerby, M. Snyder, A. Barron *et al.*, 2011 Landscape of next-generation sequencing technologies. Anal. Chem. 83(12): 4327.

Nowak, M., and R. M. May, 2000 *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, London/New York/Oxford.

Pais, D., and N. E. Leonard, 2011 Limit cycles in replicator-mutator network dynamics, edited by E. K. P. Chong, M. M. Polycarpou, J. A. Farrell, and E. F. Camacho, pp. 3922–3927 in 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC). Institute of Electrical and Electronic Engineers. Piscataway, NJ.

Park, J.-M., E. Munoz, and M. W. Deem, 2010 Quasispecies theory for finite populations. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 81(1): 011902.

Plummer, M., N. Best, K. Cowles, and K. Vines, 2006 CODA: convergence diagnosis and output analysis for MCMC. R News 6(1): 7–11.

Quiñones-Mateu, M. E., and E. J. Arts, 2002 Fitness of drug resistant HIV-1: methodology and clinical implications. Drug Resist. Updat. 5(6): 224–233.

Ramratnam, B., S. Bonhoeffer, J. Binley, A. Hurley, L. Zhang *et al.*, 1999 Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. Lancet 354 (9192): 1782–1785.

Rezende, L. F., and V. R. Prasad, 2004 Nucleoside-analog resistance mutations in HIV-1 reverse transcriptase and their influence on polymerase fidelity and viral mutation rates. Int. J. Biochem. Cell Biol. 36(9): 1716–1734.

Rodrigo, A. G., 1999 HIV evolutionary genetics. Proc. Natl. Acad. Sci. USA 96(19): 10559–10561.

Rouzine, I., and J. Coffin, 1999 Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc. Natl. Acad. Sci. USA 96(19): 10758–10763.

Schoeni-Affolter, F., B. Ledergerber, M. Rickenbach, C. Rudin, H. F. Günthard *et al.*, 2010 Cohort profile: the Swiss HIV Cohort study. Int. J. Epidemiol. 39: 1179–1189.

Segal, M. R., J. D. Barbour, and R. M. Grant, 2004 Relating HIV-1 sequence variation to replication capacity via trees and forests. Stat. Appl. Genet. Mol. Biol. 3(1): 1031.

Steinhauer, D., and J. Holland, 1987 Rapid evolution of RNA viruses. Annu. Rev. Microbiol. 41(1): 409–431.

Szendro, I. G., M. F. Schenk, J. Franke, J. Krug, and J. A. G. de Visser, 2013 Quantitative analyses of empirical fitness landscapes. J. Stat. Mech. 2013(01): P01005.

Ter Braak, C. J., 2006 A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. 16(3): 239–249.

Töpfer, A., O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin *et al.*, 2013 Probabilistic inference of viral quasispecies subject to recombination. J. Comput. Biol. 20(2): 113–123.

Vignuzzi, M., J. K. Stone, J. J. Arnold, C. E. Cameron, and R. Andino, 2005 Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439(7074): 344–348.

Wilke, C. O., 2005 Quasispecies theory in the context of population genetics. BMC Evol. Biol. 5(1): 44.

Communicating editor: J. Hermisson

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.172312/-/DC1

A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory

**David Seifert, Francesca Di Giallonardo, Karin J. Metzner,
Huldrych F. Günthard, and Niko Beerenwinkel**

Copyright © 2015 by the Genetics Society of America
DOI: 10.1534/genetics.114.172312

File S1: Supporting Information

1. Mutation matrix Q

Let C_n denote a random variable modeling a single base in generation n at some locus with state space the sequence alphabet $\mathcal{A}$. Let $c,d \in \mathcal{A}, c \neq d$, be two bases from the alphabet. The mutation rate per replication cycle is defined as the probability of not reproducing the same base

$$\mu := P\left(C_{n+1} \neq c \mid C_n = c\right) \tag{1.1}$$

As the mutation rate is assumed to be uniform for all bases, a transition from a single base to a specific other base has probability

$$P\left(C_{n+1} = d \mid C_n = c\right) = \frac{\mu}{|\mathcal{A}|-1} \tag{1.2}$$

The self-replication probability is

$$P\left(C_{n+1} = c \mid C_n = c\right) = 1-\mu \tag{1.3}$$

In order to set up the probabilities of mutation between haplotypes, we assume an independent and identical mutation rates across loci. Let $i,j \in \{1,\ldots,m\}, m = |\mathcal{A}|^L$, then we set for the mutation matrix $\mathbf{Q} = \left(q_{ij}\right)$

$$q_{ij} = \left(\frac{\mu}{|\mathcal{A}|-1}\right)^{d(i,j)} \cdot (1-\mu)^{L-d(i,j)} > 0 \tag{1.4}$$

where $d(i,j)$ denotes the Hamming distance, i.e., the number of loci at which haplotypes i and j differ. Since $q_{ij} = q_{ji}$, the matrix $\mathbf{Q}$ is symmetric.

A. Non-uniform transition/transversion rate

In order to account for a non-uniform mutation rate between different bases, the mutation model from (1.4) needs to be generalized. A mutation is called a *transition* when A ⟷ G or C ⟷ T during a replication cycle. The remaining mutations are called *transversions*, i.e., all mutations from a purine to a pyrimidine. With α we denote the probability of a transition, in line with the similar transition substitution parameter used in phylogenetic analysis. The probability of a transversion mutation occurring is denoted with β. The ratio of α/β is the transition/transversion ratio and is denoted by κ. These two mutation types can be combined to yield the overall mutation rate compatible with the definition in (1.1):

$$\mu = \alpha + 2\beta \tag{1.5}$$

The intuition of this identity is that, for every base, there exists exactly one transition mutation and two transversion mutations. The two mutation rates can now be expressed in terms of μ and κ as

$$\alpha = \mu \cdot \frac{\kappa}{\kappa+2}, \qquad \beta = \mu \cdot \frac{1}{\kappa+2} \tag{1.6}$$

For $\kappa = 1$, we find the specialization (1.2). To set up the mutation matrix for the full DNA sequence space $\mathcal{A}^L$, we use

$$q_{ij} = \alpha^{n_{\text{ti}}(i,j)} \cdot \beta^{n_{\text{tv}}(i,j)} \cdot (1-\mu)^{L-d(i,j)} \tag{1.7}$$

where $n_{\text{ti}}(i,j)$ respectively $n_{\text{tv}}(i,j)$ denote the number of transitions respectively transversions going from haplotype i to j and $d(i,j) = n_{\text{ti}}(i,j) + n_{\text{tv}}(i,j)$. It should be emphasized that, while α, β and κ bear resemblance to the parameters of the popular Kimura-2-Parameter model (also known as K80 model), the parameters used in constructing phylogenetic trees and the mutation rates here cannot be used interchangeably. Substitution parameters implicitly account for more effects, such as fixation and codon position effects, and cannot be equated with mutation rates (Kimura, 1980).

2. The function g

We ask for the equilibrium distribution $\mathbf{p} \in \Delta^{n-1}$ in the quasispecies model given a fitness landscape $\mathbf{f} \in \mathcal{F}^{n-1}$. The asterisk has been dropped from the distribution vector in (3) of the main article, as all further analysis will only be concerned with the equilibrium value of $\mathbf{p}(t)$. By (3) in the main article, for $\phi = 1$, the equilibrium distribution is

$$\mathbf{p} = \mathbf{Q}^T \operatorname{diag}(\mathbf{f})\, \mathbf{p} \tag{2.1}$$

The equilibrium distribution $\mathbf{p}$ lies in the kernel of the matrix

$$\mathbf{B} := \mathbf{Q}^T \operatorname{diag}(\mathbf{f}) - \mathbb{I}_n \tag{2.2}$$

where $\mathbb{I}_n$ denotes the $n \times n$ identity matrix. Employing the Moore-Penrose pseudoinverse (Searle, 1982), any vector in the kernel of $\mathbf{B}$ can be expressed as

$$\mathbf{a}(\mathbf{f}) := \left(\mathbb{I}_n - \mathbf{B}^+\mathbf{B}\right) \mathbb{1}_n \tag{2.3}$$

where $\mathbf{B}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{B}$ and $\mathbb{1}_n$ denotes the n-dimensional vector of all-ones. We define the scalar normalization constant $\lambda(\mathbf{f}) := \mathbb{1}_n^T \mathbf{a}(\mathbf{f})$ and set

$$g(\mathbf{f}) := \frac{\mathbf{a}(\mathbf{f})}{\lambda(\mathbf{f})} = \ker(\mathbf{B}) \cap \Delta^{n-1} \tag{2.4}$$

such that $g(\mathbf{f}) \in \Delta^{n-1}$. The function is well-defined, because $\left|\ker(\mathbf{B}) \cap \Delta^{n-1}\right| = 1$ for all $\mathbf{f} \in \mathcal{F}^{n-1}$ due to the Perron-Frobenius theorem (Bapat and Raghavan, 1997). It is not surjective, because the quasispecies equation has the property that no haplotypes can go extinct, as mutations of any haplotype will always produce all other haplotypes with non-zero probability. Thus, there exists a non-empty set of distributions, that include faces of Δ^{n-1}, which cannot arise in steady state from the quasispecies equation. We hence restrict g to its image $g : \mathcal{F}^{n-1} \to \operatorname{image}(g) =: \mathcal{Q}^{n-1} \subsetneq \Delta^{n-1}$, such that g is surjective. We refer to $\mathcal{Q}^{n-1}$ as the quasispecies space, i.e., the set of all equilibrium distributions the quasispecies equation can yield. In section B we have devised a two-haplotype model and derive lower and upper bounds on the relative frequencies defining $\mathcal{Q}^1$ that are directly related to the mutation rate of the polymerase.

A. The bijections g and h are inverses of each other

Theorem 1. *g is a bijection and h is its inverse.*

Proof. Given that g is surjective, all we have to show is

$$h(g(\mathbf{f})) = \mathbf{f} \quad \text{for all } \mathbf{f} \in \mathcal{F}^{n-1} \tag{2.5}$$

For proving (2.5), the following expansion is permissible, as $\mathbf{a}(\mathbf{f})$ is strictly positive due to the Perron-Frobenius theorem

$$\mathbf{f} = \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \operatorname{diag}(\mathbf{a}(\mathbf{f}))\, \mathbf{f} \tag{2.6}$$

$\mathbf{Q}^{-T}\mathbf{Q}^T = \mathbb{I}_n$ as $\mathbf{Q}$ is regular due to it being a strictly diagonal dominant matrix

$$\begin{aligned}
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T}\mathbf{Q}^T \operatorname{diag}(\mathbf{a}(\mathbf{f}))\, \mathbf{f} && (2.7)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T}\mathbf{Q}^T \operatorname{diag}(\mathbf{f})\, \mathbf{a}(\mathbf{f}) && (2.8)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \left(\mathbf{Q}^T \operatorname{diag}(\mathbf{f})\, \mathbf{a}(\mathbf{f}) - \mathbb{I}_n \mathbf{a}(\mathbf{f}) + \mathbf{a}(\mathbf{f})\right) && (2.9)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \left(\mathbf{B}\mathbf{a}(\mathbf{f}) + \mathbf{a}(\mathbf{f})\right) && (2.10)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \left(\mathbf{B}\left(\mathbb{I}_n - \mathbf{B}^+\mathbf{B}\right) \mathbb{1}_n + \mathbf{a}(\mathbf{f})\right) && (2.11)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \left(\left(\mathbf{B} - \mathbf{B}\mathbf{B}^+\mathbf{B}\right) \mathbb{1}_n + \mathbf{a}(\mathbf{f})\right) && (2.12)
\end{aligned}$$

We have $\mathbf{B} - \mathbf{B}\mathbf{B}^+\mathbf{B} = 0$ by definition of the Moore-Penrose pseudoinverse, hence

$$\begin{aligned}
\mathbf{f} &= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \mathbf{a}(\mathbf{f}) && (2.13)\\
&= \operatorname{diag}(\mathbf{a}(\mathbf{f}))^{-1} \mathbf{Q}^{-T} \mathbf{a}(\mathbf{f}) \frac{\lambda(\mathbf{f})}{\lambda(\mathbf{f})} && (2.14)\\
&= \operatorname{diag}\left(\frac{\mathbf{a}(\mathbf{f})}{\lambda(\mathbf{f})}\right)^{-1} \mathbf{Q}^{-T} \frac{\mathbf{a}(\mathbf{f})}{\lambda(\mathbf{f})} && (2.15)\\
&= \operatorname{diag}(g(\mathbf{f}))^{-1} \mathbf{Q}^{-T} g(\mathbf{f}) && (2.16)\\
&= h(g(\mathbf{f})) && (2.17)
\end{aligned}$$

□

B. Explicit description of $\mathcal{Q}^1$

Calculating the set $\mathcal{Q}^{n-1}$ is analytically not possible, but bounds can be formulated component-wise. Consider the two-haplotype model, where we find for the first component of $g(\mathbf{f})$, using MATLAB's symbolic toolbox,

$$p_1 = \frac{2f_2q_{21} - f_2 + 1}{f_1 - f_1q_{11} + f_2q_{21} - f_1f_2q_{11} + f_1f_2q_{21} + 1} \tag{2.18}$$

Since elements in $\mathcal{F}^{n-1}$ only have one degree of freedom when $n = 2$, we can replace f_2 with the help of the average fitness constraint $1 = p_1f_1 + p_2f_2$ and substitute into (2.18) to obtain

$$\left(\frac{1 - p_1f_1}{1 - p_1}\right)(2q_{21} - 1) + 1 = p_1\left(\left(\frac{1 - p_1f_1}{1 - p_1}\right)(q_{21} - f_1q_{11} + f_1q_{21}) + f_1 - f_1q_{11} + 1\right) \tag{2.19}$$

In the limit as $f_1 \to 0$, this equation becomes

$$0 = p_1^2 + p_1(-q_{21} - 2) + 2q_{21} \tag{2.20}$$

with roots $p_1 = q_{21}$ and $p_1 = 2$. Only the first yields a valid solution, namely $\mathbf{p} = (q_{21}, q_{22})^T$. The procedure can be repeated in an analogous fashion for $f_2 \to 0$ which then yields $\mathbf{p} = (q_{12}, q_{11})^T$. Thus, for the two-strains model, we have the component-wise bounds for $\mathbf{p} \in \mathcal{Q}^1$

$$q_{21} < p_1 < q_{11} \tag{2.21}$$

$$q_{12} < p_2 < q_{22} \tag{2.22}$$

3. Jacobian of h

In order to calculate the determinant of the Jacobian, the explicit form of the Jacobian needs to be known. Recall that

$$\operatorname{diag}(\mathbf{p})\,\mathbf{f} = \mathbf{p}\odot\mathbf{f} = \mathbf{f}\odot\mathbf{p} = \operatorname{diag}(\mathbf{f})\,\mathbf{p} \tag{3.1}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication). To determine the Jacobian of

$$h(\mathbf{p}) = \operatorname{diag}(\mathbf{p})^{-1}\mathbf{Q}^{-T}\mathbf{p}, \tag{3.2}$$

we write

$$\mathbf{p}\odot h(\mathbf{p}) = \mathbf{Q}^{-T}\mathbf{p}, \tag{3.3}$$

and perform implicit differentiation,

$$\frac{\partial}{\partial\mathbf{p}}(\mathbf{p}\odot h(\mathbf{p})) = \frac{\partial}{\partial\mathbf{p}}\left(\mathbf{Q}^{-T}\mathbf{p}\right) \tag{3.4}$$

$$\operatorname{diag}(h(\mathbf{p}))\,\mathbb{I}_n + \operatorname{diag}(\mathbf{p})\,\frac{\partial h}{\partial\mathbf{p}} = \mathbf{Q}^{-T} \tag{3.5}$$

$$\frac{\partial h}{\partial\mathbf{p}} = \operatorname{diag}(\mathbf{p})^{-1}\,\mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-1}\operatorname{diag}(h(\mathbf{p})) \tag{3.6}$$

$$\frac{\partial h}{\partial\mathbf{p}} = \operatorname{diag}(\mathbf{p})^{-1}\,\mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-1}\operatorname{diag}\left(\operatorname{diag}(\mathbf{p})^{-1}\mathbf{Q}^{-T}\mathbf{p}\right) \tag{3.7}$$

The inner-most multiplication with $\operatorname{diag}(\mathbf{p})^{-1}$ in the last term of (3.7) can be factorized as it already is a diagonal matrix, hence

$$\mathbf{J} = \frac{\partial h}{\partial\mathbf{p}} = \operatorname{diag}(\mathbf{p})^{-1}\,\mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-2}\operatorname{diag}\left(\mathbf{Q}^{-T}\mathbf{p}\right) \tag{3.8}$$

4. Functional form of posterior density function

In order to devise an efficient inference scheme, we introduce the logistic transformation (Aitchison, 1982) $t : \mathbb{R}^{n-1} \to \Delta^{n-1}$,

$$t_i(\mathbf{y}) = \begin{cases} \dfrac{\exp(y_i)}{C(\mathbf{y})} & (i = 1, \dots, n-1) \\ \dfrac{1}{C(\mathbf{y})} & (i = n) \end{cases} \qquad \mathbf{y} \in \mathbb{R}^{n-1} \tag{4.1}$$

where $C(\mathbf{y}) = 1 + \sum_{j=1}^{n-1} y_j$, and its inverse $t^{-1} : \Delta^{n-1} \to \mathbb{R}^{n-1}$,

$$t_i^{-1}(\mathbf{p}) = \log \frac{p_i}{p_n} \qquad (i = 1, \dots, n-1) \quad \mathbf{p} \in \Delta^{n-1} \tag{4.2}$$

The transformations t and t^{-1} are illustrated on the left side of Figure 1 in the main article.

We derive the functional form of the posterior density function on sample space $\mathbb{R}^{n-1}$, given data $\mathbf{X}$. This requires two transformations of the original probability density function, one from $\mathcal{F}^{n-1}$ to $\mathcal{Q}^{n-1}$ and then from $\mathcal{Q}^{n-1}$ to $\mathbb{R}^{n-1}$. For the first transformation,

$$p_{\mathcal{Q}}(\mathbf{p}) = |\det(\mathbf{J}[h](\mathbf{p}))| \cdot p_{\mathcal{F}}(h(\mathbf{p})) \tag{4.3}$$

where $p_{\mathcal{F}}(h(\mathbf{p})) = \text{const.}$ as we employ a uniform prior on $\mathcal{F}^{n-1}$

$$= |\det(\mathbf{J}[h](\mathbf{p}))| \cdot \text{const.} \tag{4.4}$$

where $p_{\mathcal{Q}}(\mathbf{p})$ denotes the transformed prior on $\mathcal{Q}^{n-1}$ and $\mathbf{J}[h](\mathbf{p}) = \dfrac{\partial h}{\partial \mathbf{p}}$ denotes the Jacobian of h with respect to $\mathbf{p}$ evaluated at some $\mathbf{p}$. We refer to section 3 for the derivation of the Jacobian

$$\mathbf{J}[h](\mathbf{p}) = \operatorname{diag}(\mathbf{p})^{-1} \mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-2} \operatorname{diag}\left(\mathbf{Q}^{-T}\mathbf{p}\right) \tag{4.5}$$

Second, we transform the previous prior on $\mathcal{Q}^{n-1}$ to $\mathbb{R}^{n-1}$. For conciseness, we calculate $\mathbf{p} = t(\mathbf{y})$ beforehand

$$\begin{align} p_{\mathbb{R}}(\mathbf{y}) &= |\det(\mathbf{J}[t](\mathbf{y}))| \cdot p_{\mathcal{Q}}(\mathbf{p} = t(\mathbf{y})) \tag{4.6} \\ &= \left(\prod_{i=1}^{n} t_i(\mathbf{y})\right) \cdot p_{\mathcal{Q}}(\mathbf{p} = t(\mathbf{y})) \tag{4.7} \end{align}$$

Substituting for $p_{\mathcal{Q}}(\mathbf{p} = t(\mathbf{y}))$ with $|\det(\mathbf{J}[h](\mathbf{p} = t(\mathbf{y})))| \cdot \text{const.}$ from (4.4) gives

$$\begin{align} &= \left(\prod_{i=1}^{n} p_i\right) \cdot |\det(\mathbf{J}[h](\mathbf{p} = t(\mathbf{y})))| \cdot \text{const.} \tag{4.8} \\ &= \left|\det\left(\mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-1} \operatorname{diag}\left(\mathbf{Q}^{-T}\mathbf{p}\right)\right)\right| \cdot \text{const.} \tag{4.9} \\ &= d(\mathbf{y}) \cdot \text{const.} \tag{4.10} \end{align}$$

where we denote the absolute value of the determinant as $d(\mathbf{y}) := \left|\det\left(\mathbf{Q}^{-T} - \operatorname{diag}(\mathbf{p})^{-1} \operatorname{diag}\left(\mathbf{Q}^{-T}\mathbf{p}\right)\right)\right|$. Thus, the posterior has the functional form

$$p_{\mathbb{R}}(\mathbf{y} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{p}) \cdot d(\mathbf{y}) \cdot \text{const.}}{P(\mathbf{X})} \tag{4.11}$$

As the normalization constant $P(\mathbf{X})$ cannot be determined, we drop it and write for the posterior density function

$$p_{\mathbb{R}}(\mathbf{y} \mid \mathbf{X}) = P(\mathbf{X} \mid \mathbf{p}) \cdot d(\mathbf{y}) = d(\mathbf{y}) \cdot \left(\prod_{i=1}^{n} p_i^{X_i}\right) \cdot \text{const.} \tag{4.12}$$

For reasons of numerical stability, we use the logarithm

$$\log p_{\mathbb{R}}(\mathbf{y} \mid \mathbf{X}) = \log d(\mathbf{y}) + \sum_{i=1}^{n} X_i \log p_i + \text{const.} \tag{4.13}$$

5. Simulations

To highlight the numerical and parameter robustness of our model, we have conducted multiple simulations. For the sake of demonstration, unless stated otherwise, we have set $\kappa = 1$.

A. Numerical precision simulations

A crucial point for numerical stability lies in calculating the determinant in $d(\mathbf{y})$ in (4.13). As a sanity check, we ran the sampling procedure with a total of 0 reads for two haplotypes, which is equivalent to sampling from the prior. A correct sampling procedure will yield a flat distribution of the random variable $f_1 - f_2$, where f_1 is the fitness of haplotype 1 and f_2 is the fitness of haplotype 2. The results are depicted in Figure S1.

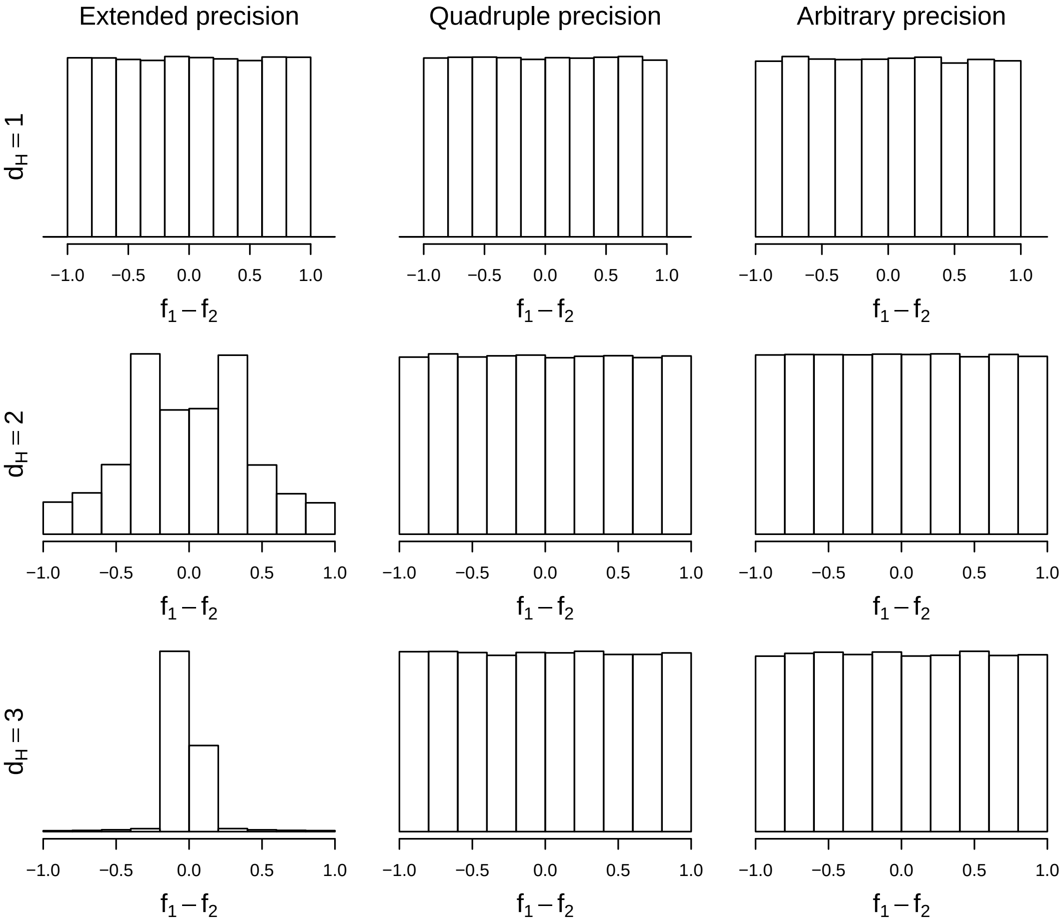

Figure S1. Prior fitness distributions for the two-haplotype model. Each column indicates a sampling procedure run with a specific precision and each row represents a haplotype constellation where haplotypes were separated by a different Hamming distance d_H.

For this simulation, the first haplotype was set to `AAA` and the second was set to `AAT`, `ATT`, and `TTT` for Hamming distances $d_H = 1$, 2, and 3, respectively. All constellations were run with $200 \cdot 10^6$ MCMC trials from the prior and a thinning interval of 1000, yielding 200 000 samples after each procedure.

The first column in Figure S1 depicts samples from the standard sampler, where floating point was performed with ordinary x87 floating point (about 18 digits of decimal precision). The second column depicts samples for 128-bit quadruple precision which was performed with GCC's `__float128` type (about 34 digits of decimal precision). The last column shows samples for running our sampling procedure with GMP's arbitrary precision type `mpf_t` (set to around 100 digits of decimal precision). Correct samplers should show a uniform distribution, as there is no fitness difference when sampling from the prior.

When the haplotype graph G_k is determined by $k = 2$, that is, the maximum number of mutations per step required for a haplotype to mutate into any other haplotype, then standard precision results cannot be trusted anymore. This is due to excessive floating-point rounding and absorption issues and motivates the requirement of $k < 2$ introduced in section *Haplotype space and mutation probabilities* of the main article. While we provide our sampler with the option of easily enabling quadruple and arbitrary precision floating point arithmetic, the performance penalties experienced by these types makes their use viable only for small haplotype sets $\mathcal{H}$.

B. Unobserved haplotypes simulations

In order to verify that the procedure detailed in the section *Haplotype space and mutation probabilities* of the main article allows for inference on data sets where the graph of observed haplotypes G_1 is not strongly connected, we conducted further simulations. We employed the same two observed haplotypes with the same varying d_H as in the previous section, that is, one observed haplotype is `AAA` and the second observed haplotype is `AAT`, `ATT` or `TTT` depending on d_H. In addition, we assumed that each haplotype was observed with exactly one read. From the symmetry of this setting and the observations, the differences of fitness values between the observed haplotypes should be symmetrical and not credibly different from 0. To circumvent the previously apparent numerical issues, we take the union of the haplotypes of the smaller d_H and the observed second haplotype for $\mathcal{H}$, such that the resulting G_1 is strongly connected. Due to the increased number of unobserved haplotypes in $\mathcal{H}$ now compared to the $\mathcal{H}$ in the previous section, the efficiency of the sampler is reduced, owing to an increased number of proposals not being an element of $\mathcal{Q}^{n-1}$. We run the sampling procedure with $100 \cdot 10^6$ MCMC trials and a thinning number of 100, the results of which are depicted in Figure S2.

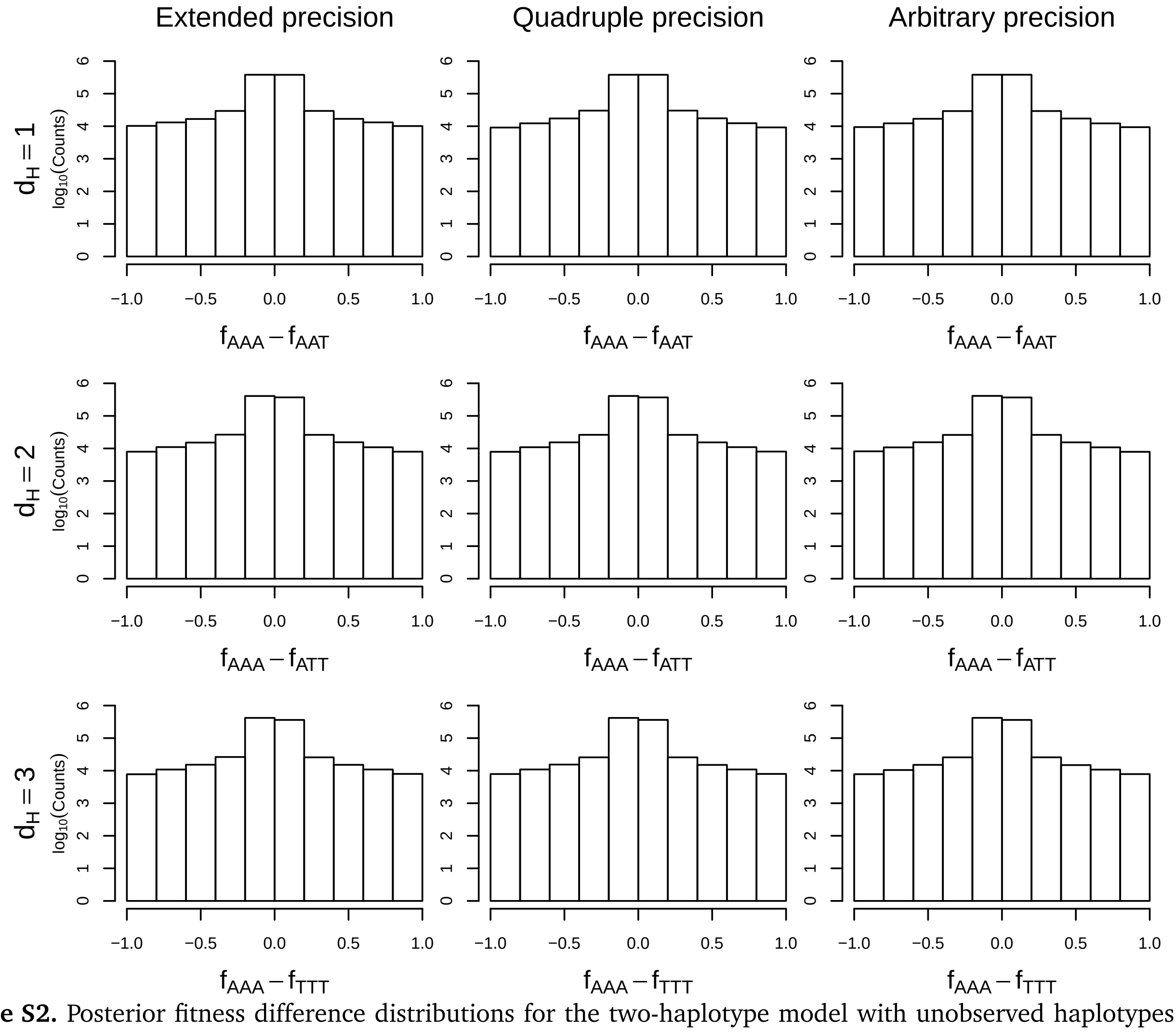

Figure S2. Posterior fitness difference distributions for the two-haplotype model with unobserved haplotypes. Each column indicates a sampling procedure run with a specific precision and each row represents a haplotype constellation where observed haplotypes were separated by a different Hamming distance d_H. As we are dealing with the posterior, the fitness differences are not uniformly distributed anymore. Due to the non-linear transformation involved in transforming probability distributions between different spaces, the tails of the posterior distribution of the fitness differences of the two observed haplotypes are heavy-tailed, hence the y-axis representing logarithmic counts.

To further assess the stability of the procedure of including unobserved haplotypes into $\mathcal{H}$, we tested whether for the same d_H, the posterior fitness samples depicted in Figure S2 come from the same distribution, i.e., whether there exists a difference between extended precision and the other numerical precision modes. To this end, we tested the difference with the Wilcoxon rank sum test, with results shown in Table S1.

As none of the differences in distributions between numerical precision modes is statistically significant at the 5% level, this demonstrates the numerical robustness of the method when including unobserved haplotypes. Lastly, as a sanity check, the 95% credibility intervals of Δf were determined for all precision modes (Table S2).

All of the credibility intervals include 0 as expected, providing a further indication that no spurious fitness differ-

Table S1. Testing for differences between precision modes for the last 50 000 samples of each run. Here $\Delta f_{\text{extended precision}}$ for instance denotes the random variable $f_{\mathtt{AAA}} - f_{\mathtt{AAT}}$ when $d_H = 1$ and extended precision was employed, i.e., the same samples as shown in Figure S2 in the top-left histogram.

$\mathbf{d_H}$	$\Delta f_{\text{extended precision}} - \Delta f_{\text{quadruple precision}}$ **p-Value**	$\Delta f_{\text{extended precision}} - \Delta f_{\text{arbitrary precision}}$ **p-Value**
1	0.3606	0.2326
2	0.1603	0.4844
3	0.6719	0.7782

Table S2. Determining 95% credibility intervals for fitness differences. All intervals include 0, such that no difference in fitness between observed haplotypes can be called.

$\mathbf{d_H}$	$\Delta f_{\text{extended precision}}$	$\Delta f_{\text{quadruple precision}}$	$\Delta f_{\text{arbitrary precision}}$
1	$[-0.614, 0.625]$	$[-0.636, 0.547]$	$[-0.600, 0.596]$
2	$[-0.563, 0.537]$	$[-0.573, 0.506]$	$[-0.558, 0.507]$
3	$[-0.548, 0.524]$	$[-0.528, 0.545]$	$[-0.530, 0.551]$

ences are called due to numerical errors.

C. Upper bound on deviation from equilibrium

To give an upper bound on how close the viral population has to be to the equilibrium, we performed dynamical simulations on the quasispecies equation. To this end, we used the same LK parameters as in the section *LK fitness landscape simulations* of the main article. The random fitness landscapes were rescaled such that the average arithmetic sum of the fitness landscape is 1. This was done to bring the average generation time to approximately one unit of time. We randomly selected one haplotype as initial starting point and simulated the system up to 10^4 time units using MATLAB. We performed the same rank-based analysis as in the *simulation studies* section of the main article, namely studying the goodness of recovering the ranks of the fitness landscape, using (6) of the main article and the ranks of the frequency vector $\mathbf{p}$. We analyzed the goodness of recovering the ranks as a function of stepping back in time, employing a total number of $N = 10\,000$ simulation points. Results for $L = 3$ are shown in Figure S3.

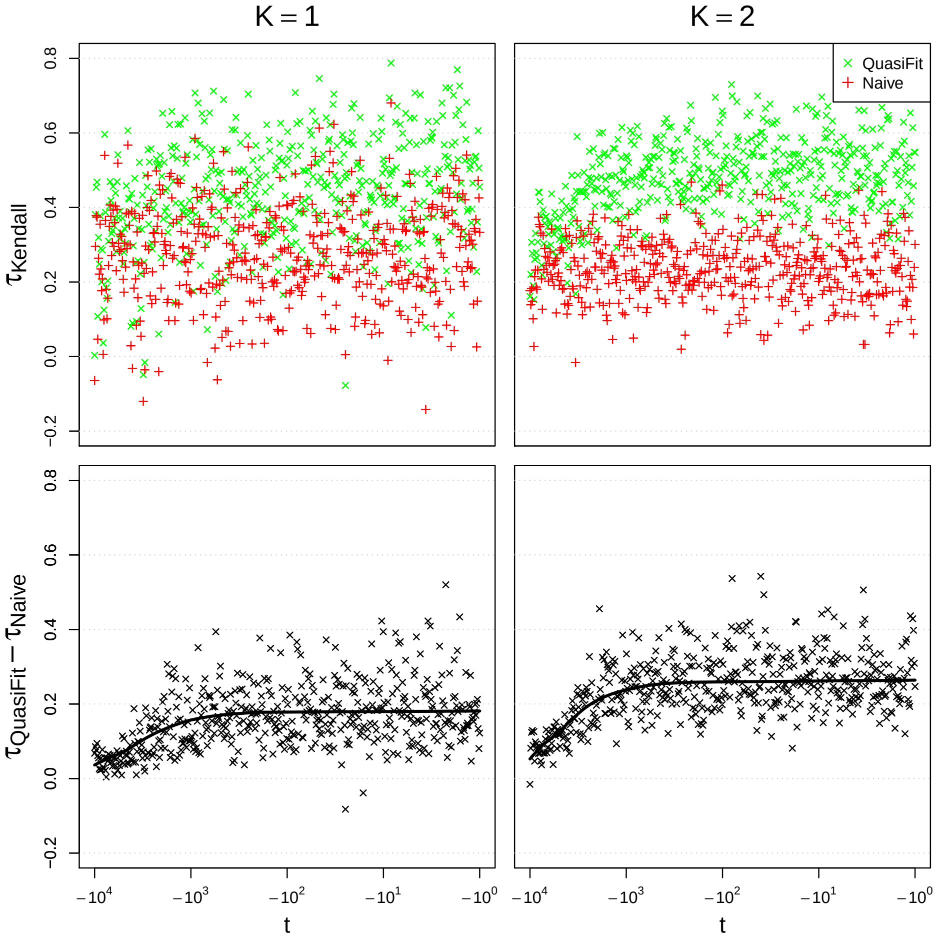

Figure S3. Accuracy of the predicted fitness landscape τ_{Kendall} as a function of the time t from equilibrium. We set $L = 3$ and analyzed the cases for $K = 1, 2$. The upper row shows the ability of the two methods to recover the fitness ranks. The bottom row illustrates the differences between the two methods. The thick solid line indicates the average distance between both methods as a function of time. For sake of clarity only 500 points are shown.

As can be seen, the *QuasiFit*-based estimator is clearly superior up to about 500 time units and degrades beyond. Nonetheless, even very far from equilibrium, the difference between both methods still marginally favors the *QuasiFit*-based estimator. Of these $N = 10\,000$ simulations, only 39 respectively 2 resulted in a better ranking of the true fitness landscape for the naive estimator for $K = 1$ respectively $K = 2$. As such, it can be concluded that the *QuasiFit*-based estimator is at least as good as the current standard of practice of taking the counts as estimator for the ranks of the fitness landscape, even when equilibrium has not been reached.

D. Deviations from transition/transversion ratio

To assess the violations of the assumed transition/transversion ratio, we conducted simulations by varying κ in (1.7). In detail, we increased κ from 1 (i.e., the uniform mutation model) up to 10 with $N = 10\,000$. For each simulation, we generated random LK fitness landscapes using the same parameters as in the previous section, calculated the quasispecies distribution using $1 < \kappa < 10$ and assumed the standard HIV mutation rate of $\mu = 3 \cdot 10^{-5}$. We then employed the standard uniform mutation matrix $\mathbf{Q}$ from (1.4) to simulate inference results for the standard *QuasiFit* case (Figure S4).

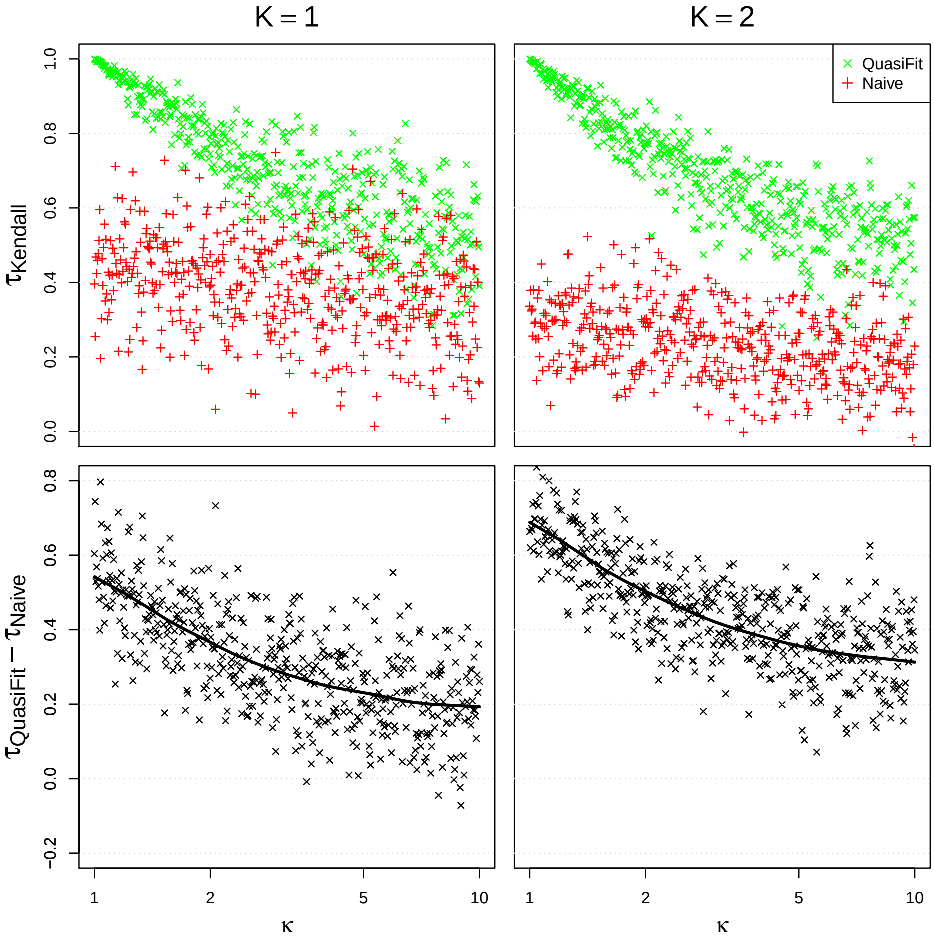

Figure S4. Accuracy of the predicted fitness landscape τ_{Kendall} as a function of the actual κ. We set $L = 3$ and analyzed the cases for $K = 1, 2$. The upper row shows the ability of the two methods to recover the fitness ranks. The bottom row illustrates the differences between the two methods. The thick solid line indicates the average distance between both methods as a function of the actual κ. For sake of clarity only 500 points are shown.

Our model is robust to at least some variation in κ. One study estimated κ to lie between 3.1 and 5.5 (Abram *et al.*, 2010). In this interval, the *QuasiFit* estimator is still better than calling fitness ranks by frequencies. In order to give the user a maximum of flexibility in inference, *QuasiFit* can also employ the mutation matrix from (1.7) to avoid possibly spurious results due to a misspecified model.

E. Epistastic vs. additive effects

In order to further understand how well the *QuasiFit* model can predict the ranks of a fitness landscape with varying levels of epistasis, we rewrite the fitness landscape as a full linear interaction model,

$$f(a_1,\ldots,a_L) = \sum_{i=1}^{L} \beta_{i,a_i} + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} \beta_{i,a_i;j,a_j} + \sum_{i=1}^{L-2}\sum_{j=i+1}^{L-1}\sum_{k=j+1}^{L} \beta_{i,a_i;j,a_j;k,a_k} + \ldots \tag{5.1}$$

where β_{i,a_i} denote the additive effects of base a at locus i, $\beta_{i,a_i;j,a_j}$ denote the pair-wise epistatic effects of base a at locus i and base b at locus j and so on. For the simulations we continued to employ the log-normal distribution as in section *LK fitness landscape simulations* of the main article. Additionally, we parametrized the log-normal distribution of the epistatic effects $\beta_{i,a_i;(\cdot)}$ such that $\mathrm{median}\left(\beta_{i,a_i;(\cdot)}/\beta_{i,a_i}\right) = C$. Hence, the epistatic and additive effects are identically distributed when $C = 1$. We refer to C as the strength of epistasis relative to the additive effects. In order for the results of this interaction model to be comparable to the results of the LK simulations, for a given K, we only included effects up to order $K+1$, e.g., if we set $K = 1$ we only included pair-wise epistatic effects $\beta_{i,a;j,b}$ and set all higher-order effects to 0.

For the simulations, we proceeded in a similar fashion as in the previous section, instead for every random fitness landscape we now generated a multinomial sample with 100 000 reads possessing a fitness MLE. Generating samples possessing an $\hat{\mathbf{f}}$ was done solely to aid inference, as $\hat{\mathbf{f}}$ can then be used as a proxy for the full Bayesian estimator. In total we simulated $N = 10\,000$ fitness landscapes with C in the interval $[3\cdot10^{-2}, 3]$. The results of the *QuasiFit* fitness rank estimator versus the naively estimated ranks are depicted in Figure S5.

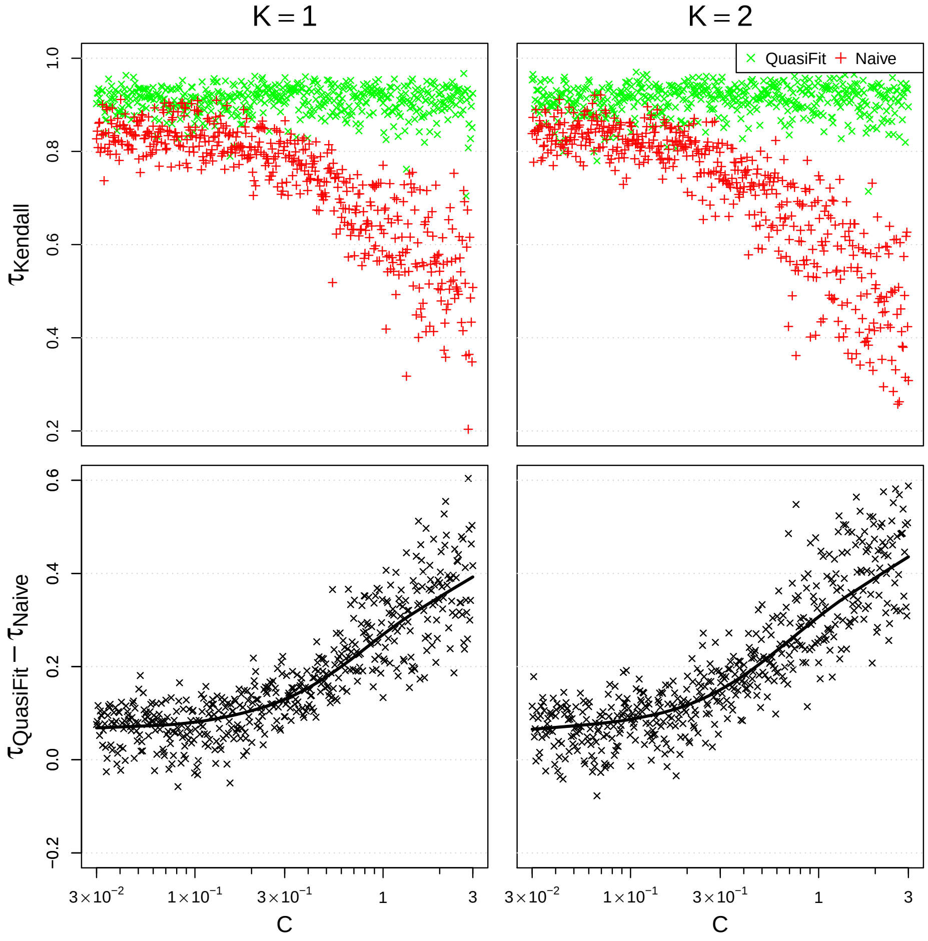

Figure S5. Accuracy of the predicted fitness landscape τ_{Kendall} as a function of the strength of epistatic relative to additive effects C. We set $L = 3$ and analyzed the cases for $K = 1,2$. The upper row shows the ability of the two methods to recover the fitness ranks. The bottom row illustrates the differences between the two methods. The thick solid line indicates the average distance between both methods as a function of the epistatic strength C. For sake of clarity only 500 points are shown.

Notice that our estimator starts to become significantly better at recovering the ranks of the fitness landscape once epistatic effects are approximately on the order of 10% of the additive effects. This detection limit can likely be decreased with increasing coverage of the reads, as the intrinsic sampling variance of the inferred fitness estimator diminishes. Assis (2014) has shown in a study of RNA secondary structure in HIV-1 that the total epistatic contribution to the fitness landscape of a locus can make up up to 50%, which is considerably larger than our lower detection limit. In addition, da Silva *et al.* (2010) have found epistasis in HIV-1 to be important and common, where the overall epistatic contribution was orders of magnitude higher than the additive contribution in several cases.

6. Convergence diagnostics

A. Gelman and Rubin diagnostic

In order to assess whether the MCMC procedure converged to its presumed stationary distribution, we analyzed the scale reduction factor for patient 1. To this end, we ran another three independent MCMC chains beside the chain on which the results reported in the main text are based. The scale factor trajectories are plotted in Figure S6.

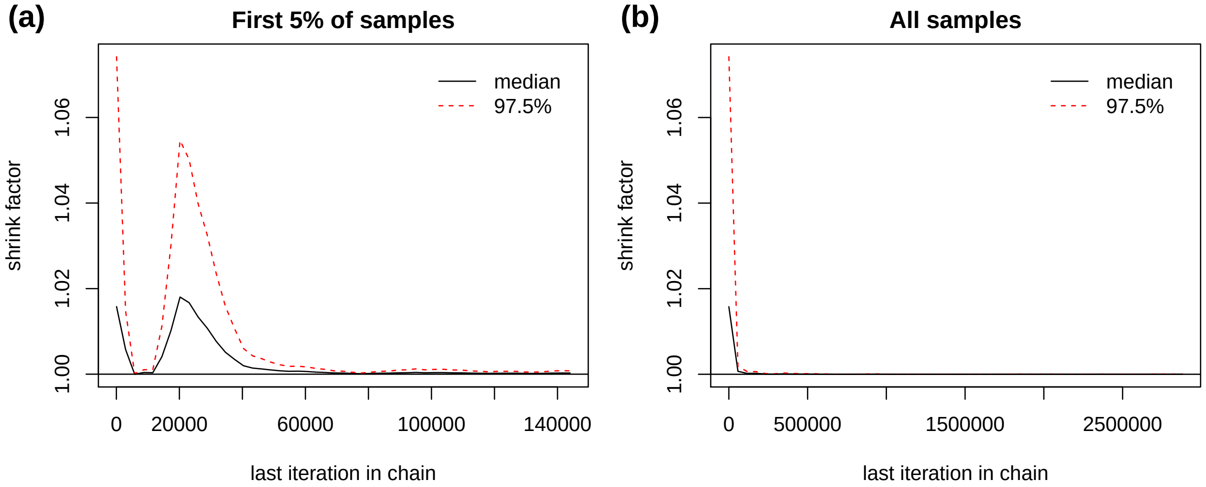

Figure S6. The Gelman-Rubin scale reduction factor. The plot in (a) shows the shrink factor vs. iteration number for the first 5% of samples. The plot in (b) shows the same shrink factor for all trial samples.

Notice how after trial count 30 000, the chains have a vanishing scale factor below 1.01, strongly suggesting convergence.

B. Autocorrelation

We determined the necessary thinning interval from autocorrelation plots (Figure S7) of one sub-chain of the MCMC procedure in the main article for patient 1.

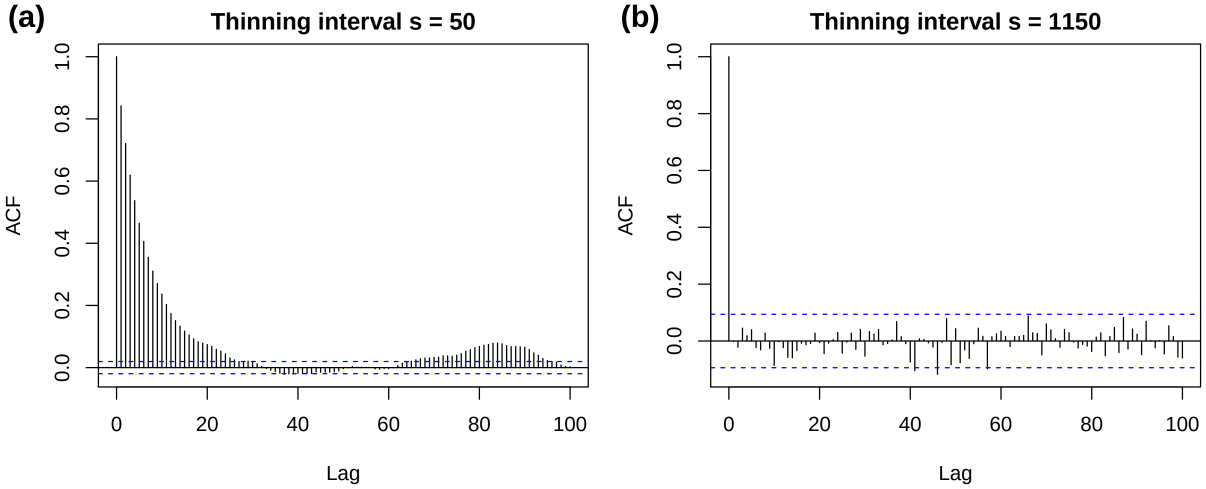

Figure S7. Autocorrelation plots for different thinning intervals. The first plot (a) indicates that even with a thinning interval of 50, significant autocorrelation remains. Plot (b) highlights that thinning interval 1150 achieves negligible autocorrelation such that samples can now be regarded as approximately independent.

At around lag 23 the autocorrelation drops below the statistical significance level. This leads to a total thinning interval of $23 \cdot 50 = 1150$ for yielding approximately independent samples from the posterior distribution.

C. Testing for differences in distributions

With thinning intervals of 1150 we proceeded to test samples from 10%–50% of trial samples with samples from 60%–100% of trial samples. Under the null hypothesis, these samples should have equal location with respect to each other if they originate from the stationary distribution. To test this null hypothesis, we employed the Wilcoxon rank sum test for all of the four independent runs in Table S3.

Table S3. Testing for differences between 40% of samples in the first half and 40% of samples in the latter half.

Runs	p-Value
1	0.3232
2	0.0751
3	0.7854
4	0.4719

None of the p-values are significant, hence we retain the null hypothesis that samples from 10%–100% originate from the same (stationary) posterior distribution.

7. Patient haplotypes

This section serves to collect the DNA sequences of haplotypes inferred from the deep sequencing data. For sake of conciseness we denote haplotypes by dropping loci with only one base and subscripting alleles at their respective loci.

Table S4. Table of haplotypes in Patient 1.

Hap. No.	Haplotype
1	$A_9A_{51}A_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
2	$A_9A_{51}A_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
3	$A_9A_{51}A_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
4	$A_9A_{51}A_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
5	$A_9A_{51}A_{74}A_{120}A_{168}C_{171}A_{183}C_{191}$
6	$A_9A_{51}A_{74}A_{120}A_{168}C_{171}G_{183}A_{191}$
7	$A_9A_{51}A_{74}A_{120}G_{168}A_{171}A_{183}A_{191}$
8	$A_9A_{51}A_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
9	$A_9A_{51}A_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
10	$A_9A_{51}A_{74}G_{120}A_{168}A_{171}A_{183}C_{191}$
11	$A_9A_{51}A_{74}G_{120}A_{168}A_{171}G_{183}A_{191}$
12	$A_9A_{51}A_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
13	$A_9A_{51}A_{74}G_{120}A_{168}C_{171}A_{183}C_{191}$
14	$A_9A_{51}A_{74}G_{120}A_{168}C_{171}G_{183}A_{191}$
15	$A_9A_{51}A_{74}G_{120}G_{168}A_{171}A_{183}A_{191}$
16	$A_9A_{51}A_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
17	$A_9A_{51}G_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
18	$A_9A_{51}G_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
19	$A_9A_{51}G_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
20	$A_9A_{51}G_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
21	$A_9A_{51}G_{74}A_{120}A_{168}C_{171}A_{183}C_{191}$
22	$A_9A_{51}G_{74}A_{120}A_{168}C_{171}G_{183}A_{191}$
23	$A_9A_{51}G_{74}A_{120}G_{168}A_{171}A_{183}A_{191}$
24	$A_9A_{51}G_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
25	$A_9A_{51}G_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
26	$A_9A_{51}G_{74}G_{120}A_{168}A_{171}A_{183}C_{191}$
27	$A_9A_{51}G_{74}G_{120}A_{168}A_{171}G_{183}A_{191}$
28	$A_9A_{51}G_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
29	$A_9A_{51}G_{74}G_{120}A_{168}C_{171}A_{183}C_{191}$
30	$A_9A_{51}G_{74}G_{120}A_{168}C_{171}G_{183}A_{191}$
31	$A_9A_{51}G_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
32	$A_9G_{51}A_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
33	$A_9G_{51}A_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
34	$A_9G_{51}A_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
35	$A_9G_{51}A_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
36	$A_9G_{51}A_{74}A_{120}A_{168}C_{171}A_{183}C_{191}$
37	$A_9G_{51}A_{74}A_{120}A_{168}C_{171}G_{183}A_{191}$
38	$A_9G_{51}A_{74}A_{120}G_{168}A_{171}A_{183}A_{191}$
39	$A_9G_{51}A_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
40	$A_9G_{51}A_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
41	$A_9G_{51}A_{74}G_{120}A_{168}A_{171}A_{183}C_{191}$
42	$A_9G_{51}A_{74}G_{120}A_{168}A_{171}G_{183}A_{191}$
43	$A_9G_{51}A_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
44	$A_9G_{51}A_{74}G_{120}A_{168}C_{171}A_{183}C_{191}$
45	$A_9G_{51}A_{74}G_{120}G_{168}A_{171}A_{183}A_{191}$
46	$A_9G_{51}A_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
47	$A_9G_{51}G_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
48	$A_9G_{51}G_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
49	$A_9G_{51}G_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
50	$A_9G_{51}G_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
51	$A_9G_{51}G_{74}A_{120}A_{168}C_{171}A_{183}C_{191}$
52	$A_9G_{51}G_{74}A_{120}A_{168}C_{171}G_{183}A_{191}$
53	$A_9G_{51}G_{74}A_{120}G_{168}A_{171}A_{183}A_{191}$
54	$A_9G_{51}G_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
55	$A_9G_{51}G_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
56	$A_9G_{51}G_{74}G_{120}A_{168}A_{171}G_{183}A_{191}$
57	$A_9G_{51}G_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
58	$A_9G_{51}G_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
59	$G_9A_{51}A_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
60	$G_9A_{51}A_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
61	$G_9A_{51}A_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
62	$G_9A_{51}A_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
63	$G_9A_{51}A_{74}A_{120}A_{168}C_{171}A_{183}C_{191}$
64	$G_9A_{51}A_{74}A_{120}A_{168}C_{171}G_{183}A_{191}$
65	$G_9A_{51}A_{74}A_{120}G_{168}A_{171}A_{183}A_{191}$
66	$G_9A_{51}A_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
67	$G_9A_{51}A_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
68	$G_9A_{51}A_{74}G_{120}A_{168}A_{171}A_{183}C_{191}$
69	$G_9A_{51}A_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
70	$G_9A_{51}A_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
71	$G_9A_{51}G_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
72	$G_9A_{51}G_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
73	$G_9A_{51}G_{74}A_{120}A_{168}A_{171}G_{183}A_{191}$
74	$G_9A_{51}G_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
75	$G_9A_{51}G_{74}A_{120}G_{168}A_{171}G_{183}A_{191}$
76	$G_9A_{51}G_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
77	$G_9A_{51}G_{74}G_{120}A_{168}C_{171}A_{183}A_{191}$
78	$G_9A_{51}G_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
79	$G_9G_{51}A_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
80	$G_9G_{51}A_{74}A_{120}A_{168}A_{171}A_{183}C_{191}$
81	$G_9G_{51}A_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
82	$G_9G_{51}A_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$
83	$G_9G_{51}A_{74}G_{120}G_{168}A_{171}G_{183}A_{191}$
84	$G_9G_{51}G_{74}A_{120}A_{168}A_{171}A_{183}A_{191}$
85	$G_9G_{51}G_{74}A_{120}A_{168}C_{171}A_{183}A_{191}$
86	$G_9G_{51}G_{74}G_{120}A_{168}A_{171}A_{183}A_{191}$

The haplotypes of Patient 1 respectively Patient 2 are noted in Table S4 respectively Table S5. The graphs of the patients' fitness landscapes are shown in Figure 9 of the main article.

Table S5. Table of haplotypes in Patient 2.

Hap. No.	Haplotype
1	$A_6A_{33}A_{72}A_{74}A_{108}G_{143}G_{144}T_{192}$
2	$A_6A_{33}A_{72}A_{74}G_{108}A_{143}G_{144}T_{192}$
3	$A_6A_{33}A_{72}A_{74}G_{108}G_{143}A_{144}C_{192}$
4	$A_6A_{33}A_{72}A_{74}G_{108}G_{143}G_{144}C_{192}$
5	$A_6A_{33}A_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
6	$A_6A_{33}A_{72}G_{74}A_{108}A_{143}G_{144}G_{192}$
7	$A_6A_{33}A_{72}G_{74}A_{108}A_{143}G_{144}T_{192}$
8	$A_6A_{33}A_{72}G_{74}A_{108}G_{143}A_{144}C_{192}$
9	$A_6A_{33}A_{72}G_{74}A_{108}G_{143}A_{144}T_{192}$
10	$A_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
11	$A_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}G_{192}$
12	$A_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
13	$A_6A_{33}A_{72}G_{74}G_{108}A_{143}A_{144}T_{192}$
14	$A_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
15	$A_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}G_{192}$
16	$A_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
17	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}C_{192}$
18	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}G_{192}$
19	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
20	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
21	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
22	$A_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
23	$A_6A_{33}G_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
24	$A_6A_{33}G_{72}G_{74}A_{108}G_{143}A_{144}T_{192}$
25	$A_6A_{33}G_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
26	$A_6A_{33}G_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
27	$A_6A_{33}G_{72}G_{74}G_{108}A_{143}A_{144}T_{192}$
28	$A_6A_{33}G_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
29	$A_6A_{33}G_{72}G_{74}G_{108}A_{143}G_{144}G_{192}$
30	$A_6A_{33}G_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
31	$A_6A_{33}G_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
32	$A_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
33	$A_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
34	$A_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
35	$A_6G_{33}A_{72}A_{74}G_{108}G_{143}G_{144}G_{192}$
36	$A_6G_{33}A_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
37	$A_6G_{33}A_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
38	$A_6G_{33}A_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
39	$A_6G_{33}A_{72}G_{74}G_{108}A_{143}G_{144}G_{192}$
40	$A_6G_{33}A_{72}G_{74}G_{108}G_{143}A_{144}G_{192}$
41	$A_6G_{33}A_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
42	$A_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
43	$A_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
44	$A_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
45	$A_6G_{33}G_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
46	$A_6G_{33}G_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
47	$A_6G_{33}G_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
48	$G_6A_{33}A_{72}A_{74}A_{108}A_{143}G_{144}T_{192}$
49	$G_6A_{33}A_{72}A_{74}A_{108}G_{143}A_{144}T_{192}$
50	$G_6A_{33}A_{72}A_{74}A_{108}G_{143}G_{144}C_{192}$
51	$G_6A_{33}A_{72}A_{74}A_{108}G_{143}G_{144}G_{192}$
52	$G_6A_{33}A_{72}A_{74}A_{108}G_{143}G_{144}T_{192}$
53	$G_6A_{33}A_{72}A_{74}G_{108}A_{143}A_{144}T_{192}$
54	$G_6A_{33}A_{72}A_{74}G_{108}A_{143}G_{144}T_{192}$
55	$G_6A_{33}A_{72}A_{74}G_{108}G_{143}A_{144}T_{192}$
56	$G_6A_{33}A_{72}A_{74}G_{108}G_{143}G_{144}C_{192}$
57	$G_6A_{33}A_{72}A_{74}G_{108}G_{143}G_{144}G_{192}$
58	$G_6A_{33}A_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
59	$G_6A_{33}A_{72}G_{74}A_{108}A_{143}A_{144}C_{192}$
60	$G_6A_{33}A_{72}G_{74}A_{108}A_{143}A_{144}T_{192}$
61	$G_6A_{33}A_{72}G_{74}A_{108}A_{143}G_{144}C_{192}$
62	$G_6A_{33}A_{72}G_{74}A_{108}A_{143}G_{144}G_{192}$
63	$G_6A_{33}A_{72}G_{74}A_{108}A_{143}G_{144}T_{192}$
64	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}A_{144}C_{192}$
65	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}A_{144}G_{192}$
66	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}A_{144}T_{192}$
67	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
68	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}G_{192}$
69	$G_6A_{33}A_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
70	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}A_{144}C_{192}$
71	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}A_{144}G_{192}$
72	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}A_{144}T_{192}$
73	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
74	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}G_{192}$
75	$G_6A_{33}A_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
76	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}C_{192}$
77	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}G_{192}$
78	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
79	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
80	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
81	$G_6A_{33}A_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
82	$G_6A_{33}G_{72}A_{74}A_{108}G_{143}G_{144}T_{192}$
83	$G_6A_{33}G_{72}A_{74}G_{108}G_{143}G_{144}C_{192}$
84	$G_6A_{33}G_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
85	$G_6A_{33}G_{72}G_{74}A_{108}A_{143}G_{144}T_{192}$
86	$G_6A_{33}G_{72}G_{74}A_{108}G_{143}A_{144}G_{192}$
87	$G_6A_{33}G_{72}G_{74}A_{108}G_{143}A_{144}T_{192}$
88	$G_6A_{33}G_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
89	$G_6A_{33}G_{72}G_{74}A_{108}G_{143}G_{144}G_{192}$
90	$G_6A_{33}G_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
91	$G_6A_{33}G_{72}G_{74}G_{108}A_{143}A_{144}T_{192}$
92	$G_6A_{33}G_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
93	$G_6A_{33}G_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
94	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}A_{144}C_{192}$
95	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}A_{144}G_{192}$
96	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
97	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
98	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
99	$G_6A_{33}G_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
100	$G_6G_{33}A_{72}A_{74}G_{108}G_{143}G_{144}T_{192}$
101	$G_6G_{33}A_{72}G_{74}A_{108}A_{143}G_{144}T_{192}$
102	$G_6G_{33}A_{72}G_{74}A_{108}G_{143}A_{144}C_{192}$
103	$G_6G_{33}A_{72}G_{74}A_{108}G_{143}A_{144}T_{192}$
104	$G_6G_{33}A_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
105	$G_6G_{33}A_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
106	$G_6G_{33}A_{72}G_{74}G_{108}A_{143}A_{144}C_{192}$
107	$G_6G_{33}A_{72}G_{74}G_{108}A_{143}A_{144}T_{192}$
108	$G_6G_{33}A_{72}G_{74}G_{108}A_{143}G_{144}C_{192}$
109	$G_6G_{33}A_{72}G_{74}G_{108}A_{143}G_{144}G_{192}$
110	$G_6G_{33}A_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
111	$G_6G_{33}A_{72}G_{74}G_{108}G_{143}A_{144}G_{192}$
112	$G_6G_{33}A_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
113	$G_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
114	$G_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
115	$G_6G_{33}A_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$
116	$G_6G_{33}G_{72}G_{74}A_{108}G_{143}G_{144}C_{192}$
117	$G_6G_{33}G_{72}G_{74}A_{108}G_{143}G_{144}T_{192}$
118	$G_6G_{33}G_{72}G_{74}G_{108}A_{143}G_{144}T_{192}$
119	$G_6G_{33}G_{72}G_{74}G_{108}G_{143}A_{144}C_{192}$
120	$G_6G_{33}G_{72}G_{74}G_{108}G_{143}A_{144}T_{192}$
121	$G_6G_{33}G_{72}G_{74}G_{108}G_{143}G_{144}C_{192}$
122	$G_6G_{33}G_{72}G_{74}G_{108}G_{143}G_{144}G_{192}$
123	$G_6G_{33}G_{72}G_{74}G_{108}G_{143}G_{144}T_{192}$

8. Codon usage effects

In the main article in section *Fitness landscapes of clinical p7 quasispecies* we analyzed the bi-allelic two loci peptide space for illustration purposes and as a proof-of-concept of our developed method. Here we show the results of looking at fitness differences of codons at synonymous loci. To this end, we iterated over all amino acid residues and analyzed those positions where heterogeneity exists in DNA sequences but not in the translated peptides. In order to analyze codon usage effects, we marginalized out the effects of all other loci, by defining equivalence classes for the synonymous codons, similar to the approach used for defining equivalence classes for peptides in section *Fitness landscapes of clinical p7 quasispecies* of the main article. We have analyzed synonymous codons for patient 1 and patient 2 and have summarized the results in Table S6 respectively Table S7.

Table S6. Codon usage in patient 1. The wild-type is indicated by the letters **wt** and defined as the major allele, whereas the mutant allele is (**mt**) defined to be the minor allele. The variable $\bar{p}$ denotes the posterior average frequency of the respective codon.

Amino acid position	**Amino acid**	**wt** Codon	$\bar{p}_{\mathrm{wt}}$	**mt** Codon	$\bar{p}_{\mathrm{mt}}$
3	Ala	`GCA`	85.5%	`GCG`	14.5%
17	Arg	`AGA`	87.8%	`AGG`	12.2%
40	Arg	`AGA`	92.1%	`AGG`	7.9%
56	Glu	`GAA`	82.5%	`GAG`	17.5%
57	Gly	`GGA`	74.6%	`GGC`	25.4%
61	Lys	`AAA`	80.0%	`AAG`	20.0%

Table S7. Codon usage in patient 2. The wild-type is indicated by the letters **wt** and defined as the major allele, whereas the mutant allele ($\mathbf{mt}_{1,2}$) is defined to be the first and (if applicable) second minor allele. The variable $\bar{p}$ denotes the posterior average frequency of the respective codon. Notice the tri-allelic locus at amino acid position 64.

Amino acid position	**Amino acid**	**wt** Codon	$\bar{p}_{\mathrm{wt}}$	$\mathbf{mt}_1$ Codon	$\bar{p}_{\mathrm{mt}_1}$	$\mathbf{mt}_2$ Codon	$\bar{p}_{\mathrm{mt}_2}$
2	Glu	`GAG`	88.4%	`GAA`	11.6%		
11	Ala	`GCA`	94.5%	`GCG`	5.5%		
24	Arg	`AGA`	91.9%	`AGG`	8.1%		
36	Gly	`GGG`	87.8%	`GGA`	12.2%		
64	Thr	`ACT`	88.4%	`ACC`	8.5%	`ACG`	3.1%

All codons could be credibly inferred to differ in their fitness, with the wild-type codon fitter than average and all mutant codons less fit than average. Given the large frequencies of the wild-type alleles, this is not unexpected. Codon usage is a known cause for fitness differences *in vivo* (Ermolaeva *et al.*, 2001).

9. Runtime evaluation

In order to better understand when the asymptotic complexity of $\mathcal{O}(n^3)$ is reached, we ran our sampler on artificial data. To this end, we reduced the alphabet to a binary set $\mathcal{A} = \{\text{A, G}\}$ and set the length of the genomic space under study to $L = \{1, \ldots, 9\}$, such that the total number of haplotypes will be $n = 2^L$. All simulations were performed with $N_{\text{trials}} = 100$ per chain and a total of 512 chains, thus having simulated a total of 51 200 MCMC trials. For each simulation, we recorded the time required for simulating the MCMC trials, divided the total runtime by 51 200 in order to yield the average runtime per MCMC trial. All simulations were conducted on an Intel Xeon E5-2697 CPU with one simulation thread. In order to estimate the transition to the asymptotic regime, we estimate two models of runtime

$$t(n) = a + b \cdot n + c \cdot n^2 + d \cdot n^3 \tag{9.1}$$

and the asymptotic model

$$t(n) = d \cdot n^3 \tag{9.2}$$

The full model (9.1) was fitted by employing non-linear least squares (NLS) on the log-transformed data, while the latter (9.2) was fitted by performing NLS on just the last three log-transformed data points. The fitted models are depicted in Figure S8 and confirm that beyond $n \approx 64$ the asymptotic regime is practically reached. In this regime the calculation of the matrix determinant in (4.13) is the rate-determining step, whereas below this limit non-cubic memory allocation and function overhead contribute a sizable portion to the computational runtime.

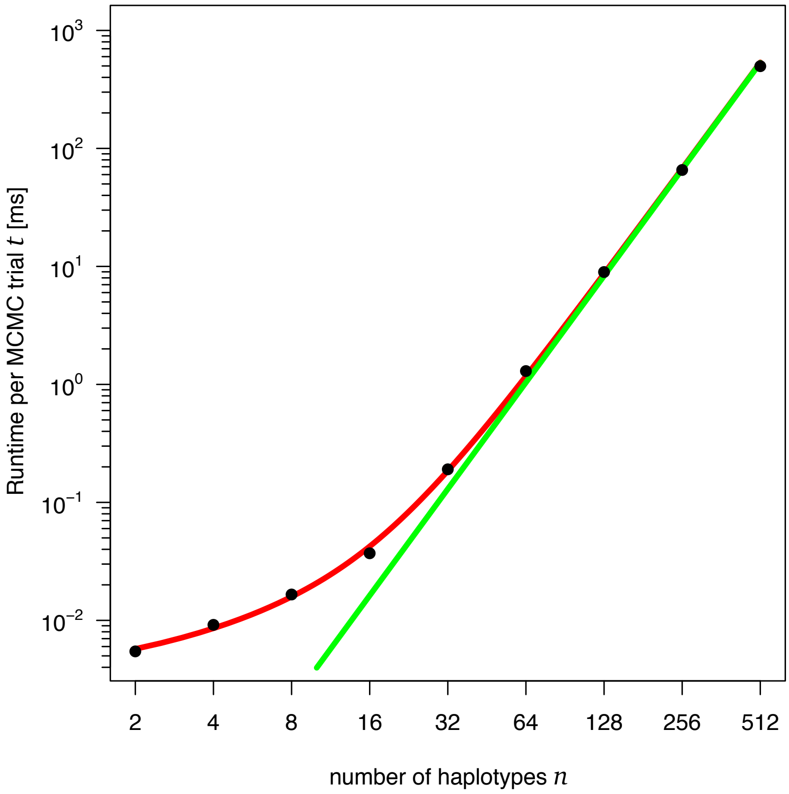

Figure S8. Graph of the per MCMC trial runtime t versus the number of haplotypes n. The red curve represents the best fit of (9.1) whereas the green model represents the asymptotic complexity (9.2).

References

Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G., and Hughes, S. H. (2010). Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of Virology*, **84**(19), 9864–9878.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(2), 139–177.

Assis, R. (2014). Strong Epistatic Selection on the RNA Secondary Structure of HIV. *PLOS Pathogens*, **10**(9), e1004363.

Bapat, R. B. and Raghavan, T. E. S. (1997). *Nonnegative matrices and applications*, volume 64. Cambridge University Press.

da Silva, J., Coetzer, M., Nedellec, R., Pastore, C., and Mosier, D. E. (2010). Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics*, **185**(1), 293–303.

Ermolaeva, M. D. *et al.* (2001). Synonymous codon usage in bacteria. *Current Issues in Molecular Biology*, **3**(4), 91–97.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**(2), 111–120.

Searle, S. (1982). *Matrix Algebra Useful for Statistics*. Wiley-Interscience.