Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Bayesian Hierarchical Modeling of the HIV Evolutionary Response to Therapy

Shane T. JENSEN, Jared PARK, Alexander F. BRAUNSTEIN, and Jon MCAULIFFE

A major challenge for the treatment of human immunodeficiency virus (HIV) infection is the development of therapy-resistant strains. We present a statistical model that quantifies the evolution of HIV populations when exposed to particular therapies. A hierarchical Bayesian approach is used to estimate differences in rates of nucleotide changes between treatment- and control-group sequences. Each group's rates are allowed to vary spatially along the HIV genome. We employ a coalescent structure to address the sequence diversity within the treatment and control HIV populations. We evaluate the model in simulations and estimate HIV evolution in two different applications: a conventional drug therapy and an antisense gene therapy. In both studies, we detect evidence of evolutionary escape response in the HIV population. Supplementary materials for this article are available online.

KEY WORDS: Coalescent; MCMC; PAC likelihood; Viral evolution.

1. INTRODUCTION

Acquired Immunodeficiency Syndrome (AIDS), the disease caused by the progression of the human immunodeficiency virus (HIV), kills almost 3 million people worldwide per year (UNAIDS 2006). Significant progress has been made toward medical therapies to decrease the mortality of the disease by slowing the progression of HIV infection. However, the HIV population within a patient can often evolve resistance to escape these treatments. An understanding of the dynamics of HIV evolution is valuable to the development of effective future therapies. In this work, we develop a statistical model to estimate the nucleotide sequence changes in the HIV genome which are involved in the evolutionary escape response to therapy.

Two common mechanisms for nucleotide sequence changes are mutation and recombination, as depicted in Figure 1. Mutation is the change in the identity of a single nucleotide at a particular location in the genome. Recombination is the exchange of larger portions of sequence between the two nonidentical copies of the HIV genome contained within each HIV virion. Recombination between these two copies can produce hybrid genomes that contain different pieces of each original sequence.

The evolutionary escape response of HIV to therapies is particularly effective in part because HIV generates between 10^9 and 10^{10} virions per day, a high replication rate relative to other viruses. HIV also has a mutation rate of 3×10^{-5} mutations per nucleotide per replication cycle, one of the highest observed in nature (Robertson et al. 1995). The consequence is high genetic diversity across the HIV population within an infected individual, which increases the possibility that a subset of the population will develop resistance to a therapy.

We model the evolutionary escape response of HIV by estimating differences in rates of mutation and recombination between two samples: genome sequences from a treatment HIV population that was exposed to a therapy versus genome sequences from a control population. Both mutation and recombination processes must be modeled simultaneously as both types of sequence change occur frequently in HIV. We must also allow spatial heterogeneity in the mutation and recombination processes, since different regions of the HIV genome may be under different evolutionary pressure due to the specific targeting of a particular therapy.

The two groups of sequences (treatment and control) are considered to be samples from different Wright–Fisher populations (Hein, Schierup, and Wiuf 2005) with their own evolutionary parameters. In the basic Wright–Fisher model, the sequence population is modeled forward in time, with sequences evolving via mutation governed by a rate parameter μ.

Kingman (1982) observed it is preferable to model the sampled sequences backward in time as they coalesce at common ancestors, forming an ancestral tree rooted at the most recent common ancestor of the sample. Kingman's resulting *coalescent* process is the continuous-time limit of this backward-looking Wright–Fisher model, where mutation events occur along the branches of the coalescent tree.

Recombination complicates matters by producing sequences that are hybrids of multiple parent sequences. Hudson (1983) generalized the coalescent process to include recombination events, which occur at a rate ρ. The genealogy of the observed sequences is no longer an ancestral tree but rather an *ancestral recombination graph*, or ARG (Griffiths and Marjoram 1996).

An ARG can be understood as a union of ancestral trees, one for each position along the sequence. For this reason, models that require us to assume that a common phylogenetic tree relates the entire sequence of the sampled individuals are potentially problematic. In fact, phylogenetic tree analyses have been shown to exhibit pronounced bias for highly recombinant populations such as viruses (Schierup and Hein 2000).

Our goal is to capture the evolutionary escape response of HIV by estimating differences in mutation and recombination rates between treatment and control populations. Ideally, we

Shane T. Jensen is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *stjensen@wharton.upenn.edu*). Jared Park is student, Department of Statistics, The University of California at Berkeley, Berkeley, CA 94720 (E-mail: *jaredjamespark@gmail.com*). Alexander F. Braunstein is student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *thestatistician@gmail.com*). Jon McAuliffe is Assistant Professor (Adjunct), Department of Statistics, The University of California at Berkeley, Berkeley, CA 94720 (E-mail: *jon@stat.berkeley.edu*). This work was supported by a grant from the Center for AIDS Research of the University of Pennsylvania. The authors thank G. Binder, B. Doms, and N. Ray for their data and advice.

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Applications and Case Studies
DOI: 10.1080/01621459.2013.830449

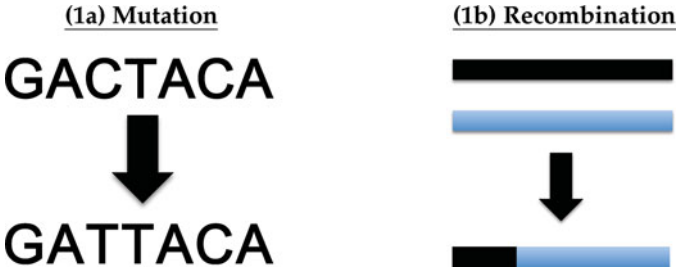

Figure 1. Pictorial representation of mechanisms for sequence change. (1a) Mutation of third nucleotide of the sequence from C to T. (1b) The black and blue lines represent two nonidentical copies of the HIV genome sequence contained within each HIV virion. When recombination occurs, a hybrid sequence is produced. The online version of this figure is in color.

would estimate these mutation and recombination parameters by integrating over all possible ARGs for our observed samples. The vast number of possible ARGs makes this calculation unrealistic.

There have been several approaches to parameter estimation in the presence of so many possible ARGs. Fearnhead and Donnelly (2001) estimated mutation and recombination rates by using importance sampling to approximately integrate over all possible ARGs. In this article, we build upon a different approach proposed by Li and Stephens (2003), where a product of approximate conditionals (PACs) scheme is used to approximately integrate over the unobserved ARG.

The PAC approach of Li and Stephens (2003) approximates the coalescent genealogy of a set of sequences by instead modeling each sequence sequentially based on the other observed sequences. Specifically, each observed sequence k is modeled as a sequence of emissions from a hidden Markov model (HMM) based on sequences $\{1, \ldots, k-1\}$.

For the lth nucleotide in sequence k, h_{kl}, we have a hidden state that specifies which sequence b in $\{1, \ldots, k-1\}$ provides the nucleotide h_{bl}. In this HMM model, the recombination rate ρ controls the transitions between hidden states, so that sequence k is potentially a hybrid of sequences $\{1, \ldots, k-1\}$. Mutations (at rate μ) allow the observed nucleotide h_{kl} to differ from the candidate nucleotide h_{bl}.

In Section 2.2, we provide additional details of this HMM approximation to the coalescent process. This PAC-HMM approach was extended by Wilson and McVean (2006) to allow spatially varying mutation and recombination rates. Similar to Wilson and McVean (2006), we allow the parameters of our evolutionary model to vary spatially along the genome, but with the additional complication that we must model potentially different sequence evolution between treatment and control populations. In addition, Wilson and McVean (2006) focused on estimating selection at the protein level, which is unsatisfactory for therapies targeting the nucleotide sequence of HIV directly. We examine one such application involving an antisense gene therapy in Section 4.2.

In Section 2, we outline a hierarchical Bayesian coalescent model that addresses several important aspects of the evolutionary escape response of HIV. The model applies to samples drawn from control and treatment HIV populations, and it is geared toward making inferences about differential effects between treatment and control. Mutation and recombination must be modeled simultaneously, as they both occur at significant rates in HIV populations. We also expect spatial heterogeneity in the mutation and recombination processes, since therapies typically target specific regions of the HIV genome.

In Section 3, we present an extensive simulation study to evaluate the ability of our model to estimate evolutionary treatment effects and underlying evolutionary parameters in different data settings. We then study the model in two different applications: a conventional drug therapy (Section 4.1) and a new antisense gene therapy (Section 4.2). In both applications, we find evidence of altered HIV evolution in response to therapeutic pressure. We compare our analysis to simpler approaches in Section 4.3 and conclude with a discussion in Section 5.

2. HIERARCHICAL BAYESIAN COALESCENT MODEL

We measure the evolutionary escape response of HIV as differences between treatment and control populations in terms of genome sequence changes induced by mutation and recombination. We employ a hierarchical prior structure to explicitly estimate these differences while still pooling information between the two populations.

The population structure within the treatment and control HIV populations is modeled using the coalescent (Kingman 1982). Our approach extends previous coalescent-based methods to allow estimation of sequence changes at the nucleotide level, where previous efforts (Wilson and McVean 2006) have focused on selection at the protein level.

Since therapies are targeted against particular regions of the HIV genome, our model also allows for spatial heterogeneity in the mutation and recombination processes. However, we still share information between proximal regions of the genome through a piecewise constant prior structure for each rate along the sequence. The model is implemented using Markov chain Monte Carlo (MCMC) techniques, which allows us to estimate the full posterior distribution of all unknown parameters.

2.1 Notation for Data and Model Parameters

The available data in both HIV applications are a set of treatment sequences $\mathbf{H}^{\mathrm{T}} = (\mathrm{h}_1^{\mathrm{T}}, \ldots, \mathrm{h}_n^{\mathrm{T}})$ and control sequences $\mathbf{H}^{\mathrm{C}} = (\mathrm{h}_1^{\mathrm{C}}, \ldots, \mathrm{h}_m^{\mathrm{C}})$. Each sequence is an aligned HIV genome consisting of five different symbols, {A,C,G,T,−}. The sequences within each set all have the same length. The fifth symbol (−) denotes a gap, since some nucleotides can be missing in a set of aligned sequences.

The generative model for the observed sequences consists of (1) a genealogy that specifies the ancestral history of the sequences, and (2) the parameters of the mutation and recombination processes that lead to sequence change. We assume that the treatment sequences $\mathbf{H}^{\mathrm{T}}$ share an unobserved genealogy $\mathbf{G}^{\mathrm{T}}$, and that the control sequences $\mathbf{H}^{\mathrm{C}}$ share an unobserved genealogy $\mathbf{G}^{\mathrm{C}}$ that is independent of $\mathbf{G}^{\mathrm{T}}$.

For the mutation process, we use the "F84" Felsenstein mutation model (Felsenstein and Churchill 1996), augmented with a gap symbol. This "F84+Gaps" model (McGuire, Denham, and Balding 2001) is controlled by two parameters: a mutation rate μ and a transition/transversion ratio κ.

The parameter μ controls the overall rate of nucleotide substitutions, which is the main parameter of interest for the mutation process. The parameter κ allows different rates of nucleotide transitions ($\mu \cdot \kappa$ for purine $\rightarrow$ purine) versus transversions (μ for purine $\rightarrow$ pyrimidine). For the recombination process, we have a single parameter ρ that specifies the rate of recombination events.

Taken together, the model thus far has a total of three parameters $\boldsymbol{\Theta} = (\mu, \kappa, \rho)$ controlling the mutation and recombination processes. We will extend this model in Section 2.3. We denote by $\boldsymbol{\Theta}^{\mathrm{T}}$ and $\boldsymbol{\Theta}^{\mathrm{C}}$ the parameters of the mutation and recombination processes for the treatment and control sequences, respectively.

2.2 Coalescent-Based PAC Likelihood

The coalescent provides the likelihood function for the observed sequence data $(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}})$ given the unknown geneologies $(\mathbf{G}^{\mathrm{T}}, \mathbf{G}^{\mathrm{C}})$ and mutation/recombination process parameters $(\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}})$. However, as discussed in Section 1, the process of recombination produces a more complicated coalescent structure called an ARG (Griffiths and Marjoram 1996).

We assume that the treatment and control populations are independent conditional on the model parameters and genealogies:

$$p(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}}, \mathbf{G}^{\mathrm{T}}, \mathbf{G}^{\mathrm{C}}) = p(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}}, \mathbf{G}^{\mathrm{T}}) \cdot p(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}}, \mathbf{G}^{\mathrm{C}}). \tag{1}$$

The ARGs $\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{C}}$ are both high-dimensional nuisance parameters that ideally would be integrated out of the likelihood:

$$p(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}}) = \sum_{\mathbf{G}^{\mathrm{T}}} p(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}}, \mathbf{G}^{\mathrm{T}}) \cdot p(\mathbf{G}^{\mathrm{T}}) \cdot \sum_{\mathbf{G}^{\mathrm{C}}} p(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}}, \mathbf{G}^{\mathrm{C}}) \cdot p(\mathbf{G}^{\mathrm{C}}), \tag{2}$$

under the additional assumption that the treatment and control ARGs are a priori independent. Conditional on particular ARGs $\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{C}}$, it is straightforward to calculate the probability distributions of the observed sequences $p(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}}, \mathbf{G}^{\mathrm{T}})$ and $p(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}}, \mathbf{G}^{\mathrm{C}})$ since all the mutation and recombination events are specified within $\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{C}}$.

However, as noted earlier, the space of possible ARGs is very large, making the direct integration in (2) impossible. Instead, we follow the PAC approach of Li and Stephens (2003), approximating the true observed-data likelihood with a PAC likelihood. This approach will also allow us to avoid the modeling of evolutionary time, which is necessary when directly estimating a coalescent model.

First consider the exact likelihood as a product of *true* conditional distributions,

$$p(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}}) = \prod_{k=1}^{n} p\left(\mathbf{H}_k^{\mathrm{T}} | \mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{T}}\right) \cdot \prod_{k=1}^{m} p\left(\mathbf{H}_k^{\mathrm{C}} | \mathbf{H}_1^{\mathrm{C}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{C}}, \boldsymbol{\Theta}^{\mathrm{C}}\right). \tag{3}$$

Here, the ARGs $\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{C}}$ have been integrated out of each conditional distribution in (3). Since these integrations cannot be done exactly, we consider PAC distributions,

$$p(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}}) \approx \underbrace{\prod_{k=1}^{n} \pi\left(\mathbf{H}_k^{\mathrm{T}} | \mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{T}}\right)}_{\mathrm{PAC}\left(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}}\right)} \cdot \underbrace{\prod_{k=1}^{m} \pi\left(\mathbf{H}_k^{\mathrm{C}} | \mathbf{H}_1^{\mathrm{C}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{C}}, \boldsymbol{\Theta}^{\mathrm{C}}\right)}_{\mathrm{PAC}\left(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}}\right)}. \tag{4}$$

Consider the calculation of $\mathrm{PAC}(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}})$ for the treatment sequences. For each approximate conditional distribution $\pi(\mathbf{H}_k^{\mathrm{T}} | \mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{T}})$, the sequence $\mathbf{H}_k^{\mathrm{T}}$ is modeled as a sequence of emissions from an HMM based on sequences $(\mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{T}})$ and parameters $\boldsymbol{\Theta}^{\mathrm{T}}$.

We denote the individual nucleotides of sequence $\mathbf{H}_k^{\mathrm{T}}$ by $(h_{k1}^{\mathrm{T}}, \ldots, h_{kl}^{\mathrm{T}}, \ldots, h_{kL}^{\mathrm{T}})$. Each observed nucleotide h_{kl}^{T} is generated based on a hidden state $b^l \in \{1, \ldots, k-1\}$. This hidden state b^l implies that the candidate nucleotide for h_{kl}^{T} is the same aligned nucleotide $h_{b^l l}^{\mathrm{T}}$ in sequence b^l.

Our mutation parameters $(\mu^{\mathrm{T}}, \kappa^{\mathrm{T}})$ control the emission probability that the observed nucleotide h_{kl}^{T} is mutated from candidate nucleotide $h_{b^l l}^{\mathrm{T}}$. Additional details for these emission calculations are given in the appendix of Wilson and McVean (2006).[1] The HMM also specifies a Markovian transition process for the hidden states $\boldsymbol{b} = (b^1, \ldots, b^l, b^{l+1}, \ldots, b^L)$, where transitions to a new underlying sequence ($b^{l+1} \neq b^l$) are controlled by our recombination parameter ρ^{T}.

In summary, the sequence $\mathbf{H}_k^{\mathrm{T}}$ is a hybrid of the $k-1$ previous sequences, with transitions between the underlying sequences dictated by the recombination parameter ρ^{T} and mutations relative to the underlying sequence dictated by the mutation parameters $(\mu^{\mathrm{T}}, \kappa^{\mathrm{T}})$. Additional discussion of this HMM model is provided by Li and Stephens (2003).[2]

With this approximate model, the usual forward summation algorithm for HMMs (Rabiner 1989) can be used to calculate the conditional distribution $\pi(\mathbf{H}_k^{\mathrm{T}} | \mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_{k-1}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{T}})$ for each k. This HMM approximation requires an ordering of the treatment sequences, so we average the calculation of $\mathrm{PAC}(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}})$ over several[3] different random orderings of the treatment sequences, as proposed by Li and Stephens (2003).

The same HMM procedure is used to calculate $\mathrm{PAC}(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}})$ for the control sequences. The result of this procedure is a likelihood that has approximately integrated out the unknown ARGs $\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{C}}$,

$$p(\mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}}) \approx \mathrm{PAC}(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\Theta}^{\mathrm{T}}) \cdot \mathrm{PAC}(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\Theta}^{\mathrm{C}}). \tag{5}$$

[1] The Wilson and McVean (2006) calculations have mutation at the codon level, but can be adapted to our model with mutation on the nucleotide level.

[2] We use the PAC-A version of the approach given by Li and Stephens (2003).

[3] We use 10 orderings in our applications in Section 4, though our simulation experiments suggested that even five orderings are sufficient for a stable estimate of the PAC likelihood.

2.3 Prior Structure for Mutation/Recombination Parameters

As discussed in Section 2.1, the parameters of our model for sequence change consist of a recombination rate ρ as well as mutation rate μ and transition–transversion ratio κ. The last two together dictate the probability of substitutions between the five symbols $\{A, C, G, T, -\}$. We model the evolutionary escape response of HIV by allowing the recombination and mutation rate parameters to differ between the treatment and control populations, while sharing information between the two groups through a common prior distribution. Since these rates are all strictly positive, a reasonable choice of prior is for the natural logarithm of the mutation rates $(\log \mu^{\mathrm{T}}, \log \mu^{\mathrm{C}})$ to share a common normal distribution, and similarly for the recombination rates $(\log \rho^{\mathrm{T}}, \log \rho^{\mathrm{C}})$. We also share information between the two populations by assuming a common transition–transversion ratio κ.

However, since most therapies are expected to target a specific region of the HIV genome, we must also model heterogeneity along the genomic sequence. Specifically, we allow the mutation and recombination rates to vary spatially by using a piecewise constant blocking structure for the spatial variation prior. This prior allows us to share information along the genome (as well as between treatment and control populations).

Consider first the mutation rate μ for a single population of HIV genomic sequences, each of length L. We partition the HIV genome into B_μ blocks and allow each block i to have its own mutation rate μ_i. We use $\boldsymbol{\mu}$ to denote the set of mutation rates across all B_μ blocks. The number of blocks B_μ is random, with prior distribution

$$B_\mu \sim \text{Binomial}(L - 2, \delta_\mu), \tag{6}$$

where δ_μ is a fixed hyperparameter that specifies a prior expectation on the number of blocks. In our two applications (Sections 4.1–4.2), we used different values of δ_μ that implied an expected number of blocks between 30 and 60, and found that posterior inference was not sensitive to these hyperparameter values.[4] The unknown locations of the B_μ breakpoints are assumed a priori to be uniformly distributed across the length of the genome.

We share information between the mutation rates μ_i in each block i through a common prior distribution,

$$\log \mu_i \sim \text{Normal}\big(\log \mu_0, \sigma_\mu^2\big), \quad i = 1, \ldots, B_\mu, \tag{7}$$

where μ_0 is a central mutation rate for the entire HIV genome, and σ_μ^2 is the between-block variance in mutation rates. Within each block i, we allow the mutation rates to vary between the treatment and control populations, while once more sharing information with a common prior distribution,

$$\big(\log \mu_i^{\mathrm{T}}, \log \mu_i^{\mathrm{C}}\big) \overset{\text{iid}}{\sim} \text{Normal}\big(\log \mu_i, \tau_\mu^2\big), \quad i = 1, \ldots, B_\mu. \tag{8}$$

Here τ_μ^2 is the within-block variance between the treatment and control populations. The global parameters μ_0, σ_μ^2, and τ_μ^2 are all unknown. We place a normal prior on μ_0 and inverse-Gamma priors on σ_μ^2 and τ_μ^2. We use $\boldsymbol{\mu}^{\mathrm{T}}$ and $\boldsymbol{\mu}^{\mathrm{C}}$ to denote the set of treatment and control mutation rates across all B_μ blocks.

[4]Our inference in simulation studies in Section 3.1 was also not sensitive to the values of δ_μ and δ_ρ.

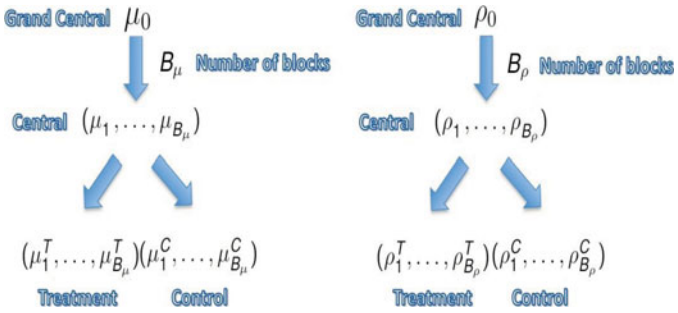

Figure 2. Hierarchical structure of prior distribution for spatially varying mutation rates $\boldsymbol{\mu}^{\mathrm{T}}$ and $\boldsymbol{\mu}^{\mathrm{C}}$ (left) and recombination rates $\boldsymbol{\rho}^{\mathrm{T}}$ and $\boldsymbol{\rho}^{\mathrm{C}}$ (right). The online version of this figure is in color.

The hierarchical structure of our prior distribution for the spatially varying mutation rates $\boldsymbol{\mu}^{\mathrm{T}}$ and $\boldsymbol{\mu}^{\mathrm{C}}$ is illustrated on the left-hand side of Figure 2. The right-hand side of Figure 2 shows a similar structure for spatially varying recombination rates $\boldsymbol{\rho}^{\mathrm{T}}$ and $\boldsymbol{\rho}^{\mathrm{C}}$. For recombination, we break the HIV genome into B_ρ blocks, where each block has a different recombination rate ρ_i with a common normal prior centered at ρ_0 and having variance σ_ρ^2. Each ρ_i serves as the center of a normal distribution (with variance τ_ρ^2) for the treatment and control recombination rates ρ_i^{T} and ρ_i^{C}. Prior distributions for the global recombination parameters ρ_0, σ_ρ^2, and τ_ρ^2 are analogous to the mutation-rate case.

Wilson and McVean (2006) also employed a block prior structure for evolutionary model parameters, but without a hierarchical structure between treatment and control populations. Also, their model focused on selection at the protein level, whereas we model mutation at the nucleotide level.

Finally, we specify an exponential prior distribution for the global transition–transversion ratio, $\kappa \sim \text{Exp}(\kappa_0)$, where κ_0 is a hyperparameter set to 0.3 in our two applications (other values of κ_0 did not materially alter the results). The prior distributions in Equations (6)–(8) for the mutation and recombination rates are combined with the PAC likelihood in Equation (5) to give the posterior distribution,

$$\begin{aligned} &p(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{C}}, \boldsymbol{\rho}^{\mathrm{T}}, \boldsymbol{\rho}^{\mathrm{C}}, \kappa, \boldsymbol{\Delta} \mid \mathbf{H}^{\mathrm{T}}, \mathbf{H}^{\mathrm{C}}) \\ &\quad \propto \text{PAC}(\mathbf{H}^{\mathrm{T}} | \boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\rho}^{\mathrm{T}}, \kappa) \cdot \text{PAC}(\mathbf{H}^{\mathrm{C}} | \boldsymbol{\mu}^{\mathrm{C}}, \boldsymbol{\rho}^{\mathrm{C}}, \kappa) \\ &\qquad \cdot p(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{C}}, \boldsymbol{\rho}^{\mathrm{T}}, \boldsymbol{\rho}^{\mathrm{C}} | \boldsymbol{\Delta}) \cdot p(\boldsymbol{\Delta}) \cdot p(\kappa). \end{aligned} \tag{9}$$

Here $\boldsymbol{\Delta} = (B_\mu, B_\rho, \boldsymbol{\mu}, \boldsymbol{\rho}, \mu_0, \rho_0, \sigma_\mu^2, \sigma_\rho^2, \tau_\mu^2, \tau_\rho^2)$ collects the additional parameters of our hierarchical prior distribution. We estimate this posterior distribution using MCMC simulation as outlined in the next section.

2.4 Markov Chain Monte Carlo Implementation

We approximate the posterior distribution (9) with a sampling scheme that combines Gibbs (Geman and Geman 1984), Metropolis–Hastings (Hastings 1970), and reversible jump (Green 1995) steps. Below, we outline the specific steps for sampling the mutation process parameters. The same steps are used *mutatis mutandis* for sampling the parameters of the recombination process parameters.

2.4.1 Sampling Changes to Blocking Structure of $(\boldsymbol{\mu}, \boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{C}})$. We change the number of blocks B_μ by proposing to

either split an existing block into two blocks or merge two adjacent blocks into a single block. We choose to split (vs. merge) with probability $s(B_\mu)/(s(B_\mu)+m(B_\mu))$, where

$$s(B_\mu) = \min\left(1, \frac{p(B_\mu+1)}{p(B_\mu)}\right) \quad \text{and}$$
$$m(B_\mu) = \min\left(1, \frac{p(B_\mu-1)}{p(B_\mu)}\right). \tag{10}$$

In both equations in Equation (10), $p(x)$ is the probability of x blocks under the binomial prior (6). If a split choice is made, a block j (with mutation rates μ_j, μ_j^{T}, and μ_j^{C}) is randomly chosen and $U \sim \text{Uniform}(0,1)$ is also sampled. New blocks j and $j+1$ are then proposed with central mutation rates $\mu_j^\star$ and $\mu_{j+1}^\star$, where $\mu_j^\star \mu_{j+1}^\star = \mu_j$ and $\mu_{j+1}^\star/\mu_j^\star = (1-U)/U$. New treatment mutation rates $\mu_j^{\mathrm{T}\star}$ and $\mu_{j+1}^{\mathrm{T}\star}$ are also proposed, where $\mu_j^{\mathrm{T}\star}\mu_{j+1}^{\mathrm{T}\star} = \mu_j^{\mathrm{T}}$ and $\mu_{j+1}^{\mathrm{T}\star}/\mu_j^{\mathrm{T}\star} = (1-U)/U$. A similar proposal is made for control mutation rates.

The split proposal is accepted with probability $\min\{R,1\}$, where

$$R = \frac{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}\star})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}\star})}{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}})}\frac{\delta_\mu}{1-\delta_\mu}\frac{s(B_\mu+1)}{m(B_\mu)} \times \frac{L-B_\mu-1}{B_\mu+1} r_\mu r_{\mu\mathrm{T}} r_{\mu\mathrm{C}}. \tag{11}$$

The vector $(\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}})$ contains the current set of parameters, whereas $(\boldsymbol{\Theta}^{\mathrm{T}\star}, \boldsymbol{\Theta}^{\mathrm{C}\star})$ has the proposed $\boldsymbol{\mu}^{\mathrm{T}\star}$ and $\boldsymbol{\mu}^{\mathrm{C}\star}$ that contain the split move. The value r_μ equals

$$\frac{\exp\left(-((\log\mu_j^\star - \log\mu_0)^2 + (\log\mu_{j+1}^\star - \log\mu_0)^2)/2\sigma_\mu^2\right)}{\exp\left(-(\log\mu_j - \log\mu_0)^2/2\sigma_\mu^2\right)} \times \frac{(\mu_j^\star+\mu_{j+1}^\star)^2}{\mu_j^\star\mu_{j+1}^\star\sigma_\mu\sqrt{2\pi}}. \tag{12}$$

The values $r_{\mu\mathrm{T}}$ (or $r_{\mu\mathrm{C}}$) have the same form as r_μ except that each μ_j value is replaced with the corresponding μ_j^{T} (or μ_j^{C}) values, μ_0 is replaced with μ_j, and σ_μ is replaced with τ_μ.

If a merge choice is made, two neighboring blocks j and $j+1$ (with mutation rates $\mu_j, \mu_{j+1}, \mu_j^{\mathrm{T}}, \mu_{j+1}^{\mathrm{T}}, \mu_j^{\mathrm{C}}, \mu_{j+1}^{\mathrm{C}}$) are randomly chosen and $U \sim \text{Uniform}(0,1)$ is also sampled. We propose to merge these blocks into one block with central mutation rate $\mu_j^\star = \mu_j\mu_{j+1}$, treatment mutation rate $\mu_j^{\mathrm{T}\star} = \mu_j^{\mathrm{T}}\mu_{j+1}^{\mathrm{T}}$, and control mutation rate $\mu_j^{\mathrm{C}\star} = \mu_j^{\mathrm{C}}\mu_{j+1}^{\mathrm{C}}$.

The merge proposal is accepted with probability $\min\{Q,1\}$ where

$$Q = \frac{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}\star})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}\star})}{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}})}\frac{1-\delta_\mu}{\delta_\mu}\frac{m(B_\mu-1)}{s(B_\mu)} \times \frac{B_\mu}{L-B_\mu} q_\mu q_{\mu\mathrm{T}} q_{\mu\mathrm{C}}. \tag{13}$$

The vector $(\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}})$ contains the current set of parameters, whereas $(\boldsymbol{\Theta}^{\mathrm{T}\star}, \boldsymbol{\Theta}^{\mathrm{C}\star})$ has the proposed $\boldsymbol{\mu}^{\mathrm{T}\star}$ and $\boldsymbol{\mu}^{\mathrm{C}\star}$ that contain the merge move. The value q_μ equals

$$\frac{\exp\left(-(\log\mu_j^\star - \log\mu_0)^2/2\sigma_\mu^2\right)}{\exp\left(-((\log\mu_j - \log\mu_0)^2 + (\log\mu_{j+1} - \log\mu_0)^2)/2\sigma_\mu^2\right)} \times \frac{\mu_j\mu_{j+1}\sigma_\mu\sqrt{2\pi}}{(\mu_j+\mu_{j+1})^2}. \tag{14}$$

The values $q_{\mu\mathrm{T}}$ (or $q_{\mu\mathrm{C}}$) have the same form as q_μ, except that each μ_j variable is replaced with the corresponding μ_j^{T} (or μ_j^{C}) variables, μ_0 is replaced with μ_j, and σ_μ is replaced with τ_μ.

We also consider changes to the blocking structure that shift the boundaries of a particular block while retaining the same number of blocks B_μ. We choose one of the $B_\mu - 1$ moveable (internal) boundaries at random and draw a value x from a geometric distribution with mean $v = 5$. We then propose to shift the chosen boundary x positions to the left or right (with equal probability). If the proposed shift crosses another boundary or creates a block of size less than two, it is automatically rejected. Otherwise, we accept the shift with probability $\min\{S,1\}$, where

$$S = \frac{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}\star})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}\star})}{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}})}. \tag{15}$$

The vector $(\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}})$ contains the current blocking structure, while $(\boldsymbol{\Theta}^{\mathrm{T}\star}, \boldsymbol{\Theta}^{\mathrm{C}\star})$ contains the proposed shift in the blocking structure.

2.4.2 Sampling Treatment, Control, and Central Mutation Rates $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{C}}, \boldsymbol{\mu})$. For each block j, we sample new values of the treatment and control mutation rates $(\mu_j^{\mathrm{T}}, \mu_j^{\mathrm{C}})$ by first proposing new values

$$\log\mu_j^{\mathrm{T}\star} \sim \text{Normal}\big(\log\mu_j^{\mathrm{T}}, \gamma^2\big) \quad \text{and}$$
$$\log\mu_j^{\mathrm{C}\star} \sim \text{Normal}\big(\log\mu_j^{\mathrm{C}}, \gamma^2\big), \tag{16}$$

and then accepting these values with probability $\min\{1,P\}$ where

$$P = \frac{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}\star})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}\star})}{\text{PAC}(\mathbf{H}^{\mathrm{T}}|\boldsymbol{\Theta}^{\mathrm{T}})\cdot\text{PAC}(\mathbf{H}^{\mathrm{C}}|\boldsymbol{\Theta}^{\mathrm{C}})} \times \frac{\mu_j^{\mathrm{T}}\mu_j^{\mathrm{C}}}{\mu_j^{\mathrm{T}\star}\mu_j^{\mathrm{C}\star}} \times \frac{\exp\left(-\left(\left(\log\mu_j^{\mathrm{T}\star} - \log\mu_j\right)^2 + \left(\log\mu_j^{\mathrm{C}\star} - \log\mu_j\right)^2\right)/2\tau_\mu^2\right)}{\exp\left(-\left(\left(\log\mu_j^{\mathrm{T}} - \log\mu_j\right)^2 + \left(\log\mu_j^{\mathrm{C}} - \log\mu_j\right)^2\right)/2\tau_\mu^2\right)}. \tag{17}$$

The vector $(\boldsymbol{\Theta}^{\mathrm{T}}, \boldsymbol{\Theta}^{\mathrm{C}})$ contains the current set of parameters whereas $(\boldsymbol{\Theta}^{\mathrm{T}\star}, \boldsymbol{\Theta}^{\mathrm{C}\star})$ has the proposed $\mu_j^{\mathrm{T}\star}$ and $\mu_j^{\mathrm{C}\star}$. The variance parameter γ^2 of the proposed distributions was tuned adaptively to achieve good mixing properties of the Markov chain.

The central mutation rates $\boldsymbol{\mu}$ can be sampled directly without needing Metropolis–Hastings steps. For each block j, we sample μ_j as follows:

$$\log\mu_j \sim \text{Normal}\left(\frac{\frac{\log\mu_j^{\mathrm{T}} + \log\mu_j^{\mathrm{C}}}{\tau_\mu^2} + \frac{B_\mu}{\sigma_\mu^2}\log\mu_0}{\frac{2}{\tau_\mu^2} + \frac{B_\mu}{\sigma_\mu^2}}, \frac{1}{\frac{2}{\tau_\mu^2} + \frac{B_\mu}{\sigma_\mu^2}}\right). \tag{18}$$

2.4.3 Sampling Global Parameters $(\mu_0, \sigma_\mu^2, \tau_\mu^2, \kappa)$. The mutation process global parameters μ_0, σ_μ^2, and τ_μ^2 all have

standard conditional posterior distributions:

$$\log \mu_0 \sim \text{Normal}\left(\frac{\sum_j \log \mu_j}{B_\mu}, \frac{\sigma_\mu^2}{B_\mu}\right),$$

$$\left(\sigma_\mu^2\right)^{-1} \sim \text{Gamma}\left(\frac{B_\mu}{2}+1, \frac{1}{2}\left[\sum_j (\log \mu_j - \log \mu_0)^2\right]\right),$$

$$\left(\tau_\mu^2\right)^{-1} \sim \text{Gamma}\left(B_\mu + 1, \frac{1}{2}\left[\sum_j \left(\log \mu_j^{\text{T}} - \log \mu_j\right)^2 + \left(\log \mu_j^{\text{C}} - \log \mu_j\right)^2\right]\right).$$

The transition–transversion ratio κ is sampled by considering a proposal value $\kappa^\star = \exp(U)\kappa$, where $U \sim \text{Uniform}(-1, 1)$, and accepting $\kappa^\star$ with probability $\min\{1, K\}$, where

$$K = \frac{\text{PAC}(\mathbf{H}^{\text{T}}|\boldsymbol{\Theta}^{\text{T}\star}) \cdot \text{PAC}(\mathbf{H}^{\text{C}}|\boldsymbol{\Theta}^{\text{C}\star})}{\text{PAC}(\mathbf{H}^{\text{T}}|\boldsymbol{\Theta}^{\text{T}}) \cdot \text{PAC}(\mathbf{H}^{\text{C}}|\boldsymbol{\Theta}^{\text{C}})} \frac{\kappa^\star}{\kappa} \frac{\exp(-\kappa_0\kappa^\star)}{\exp(-\kappa_0\kappa)}.$$

The vector $(\boldsymbol{\Theta}^{\text{T}}, \boldsymbol{\Theta}^{\text{C}})$ contains the current set of parameters, whereas $(\boldsymbol{\Theta}^{\text{T}\star}, \boldsymbol{\Theta}^{\text{C}\star})$ contains the proposed $\kappa^\star$.

2.4.4 Convergence and Running Time. Eight chains were run for several hundred thousand iterations to ensure convergence to sampling from the posterior distribution (9). Post-convergence chains were combined and thinned to eliminate autocorrelation. Our MCMC algorithm is computationally intensive, with each chain running for two to three days on a cluster. The rate-limiting calculations are the PAC likelihoods $\text{PAC}(\mathbf{H}^{\text{T}}|\boldsymbol{\Theta}^{\text{T}})$ and $\text{PAC}(\mathbf{H}^{\text{C}}|\boldsymbol{\Theta}^{\text{C}})$, as outlined in Section 2.2. These PAC-likelihood evaluations must be performed for each acceptance decision in Sections 2.4.1–2.4.3.

3. SIMULATION STUDIES OF OUR APPROACH

In Section 3.1, we evaluate our ability to estimate *treatment effects*: differences in mutation and recombination between treatment and control sequences, which are the primary focus of our real data analyses. Specifically, we evaluate the posterior coverage for the differential mutation rates $(\boldsymbol{\mu}^{\text{T}} - \boldsymbol{\mu}^{\text{C}})$ and differential recombination rates $(\boldsymbol{\rho}^{\text{T}} - \boldsymbol{\rho}^{\text{C}})$ between treatment and control sequence datasets that were generated under simple conditions outlined in Section 3.1.

In Section 3.2, we evaluate our ability to estimate the underlying parameters of our model beyond treatment effects. Specifically, we examine coverage rates and estimation accuracy for both global parameters (e.g., μ_0, ρ_0, κ) and within-block parameters (e.g., $\boldsymbol{\mu}, \boldsymbol{\rho}$) under a variety of data settings, including conditions that are intended to mimic the real data applications of Section 4.

3.1 Simulation Evaluation: Coverage of Treatment Effect

We use the coalescent simulator MS (Hudson 2002) with SEQ-GEN (Rambaut and Grassly 1997) to generate treatment and control sequences from the coalescent with known spatially varying mutation and recombination rates. In the first experiment, we generate sequence datasets where there is a true spatially varying treatment effect in the mutation rate $(\boldsymbol{\mu}^{\text{T}} - \boldsymbol{\mu}^{\text{C}} \neq \mathbf{0})$ but no treatment effect in terms of recombination $(\boldsymbol{\rho}^{\text{T}} - \boldsymbol{\rho}^{\text{C}} = \mathbf{0})$. In the second experiment, there is not only a true spatially varying treatment effect in the mutation rate $(\boldsymbol{\mu}^{\text{T}} - \boldsymbol{\mu}^{\text{C}} \neq \mathbf{0})$ but also a true constant treatment effect in the recombination rate $(\boldsymbol{\rho}^{\text{T}} - \boldsymbol{\rho}^{\text{C}} = \boldsymbol{c} \neq \mathbf{0})$. The true mutation and recombination treatment effects for these two experiments are shown in Figure 3.

The simulated treatment and control populations each consisted of 20 sequences, with each sequence 600 nucleotides in

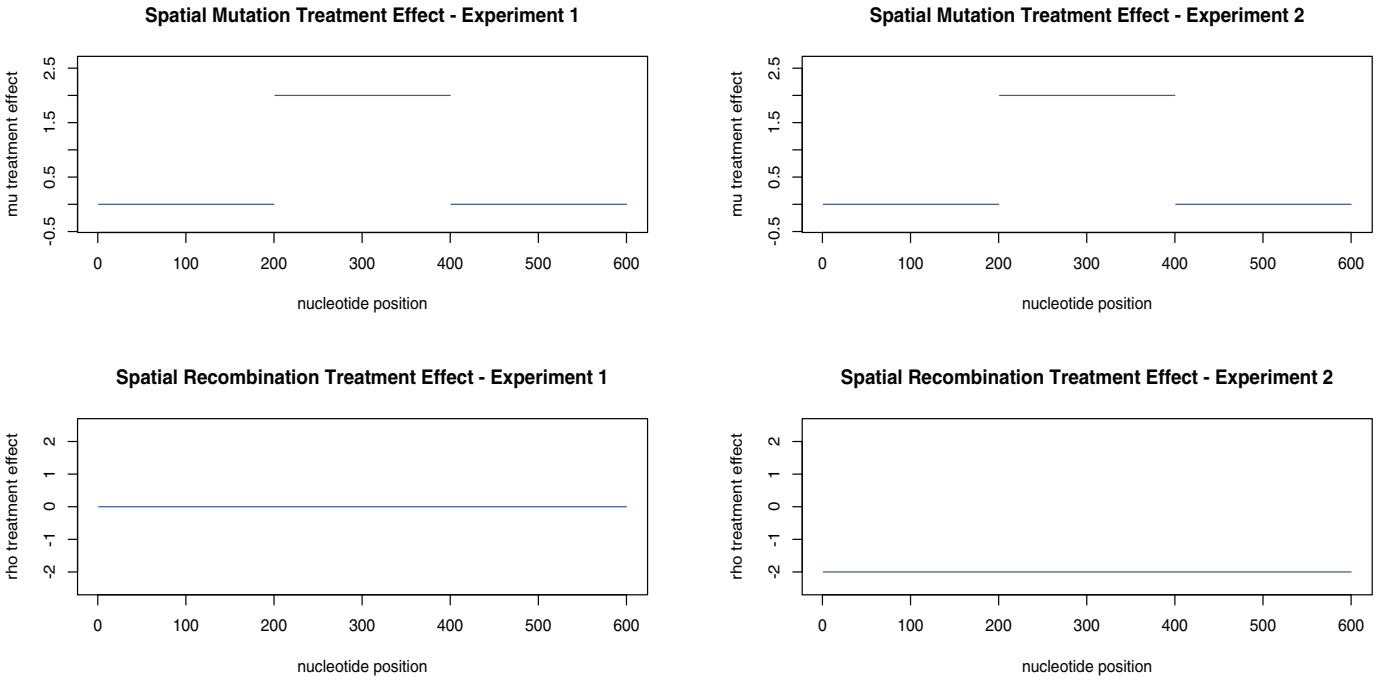

Figure 3. Left: Experiment 1 has a spatially varying mutation treatment effect but no recombination treatment effect. Right: Experiment 2 has a spatially varying mutation treatment effect and a constant but nonzero recombination treatment effect. The online version of this figure is in color.

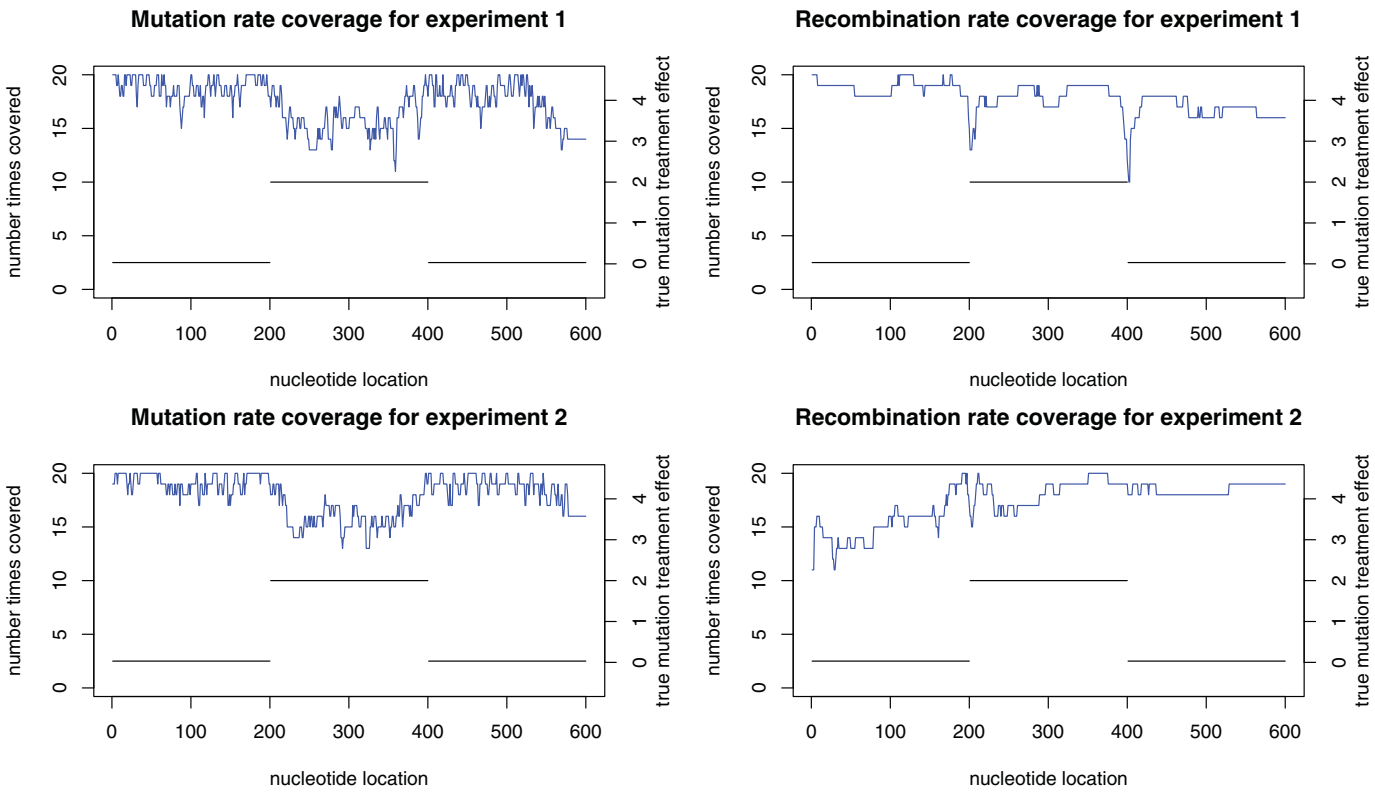

Figure 4. Coverage of true mutation and recombination treatment effects for Experiments 1 and 2. Blue lines indicate how many times (out of 20 replicates) the 95% HPD interval covered the true treatment effect. Black lines give the true mutation treatment effect. The online version of this figure is in color.

length. The number and length of these simulated sequences is similar to the applications in Sections 4.1 and 4.2. Both experiments were replicated 20 times.

We focus our analysis on the coverage of the 95% highest-posterior-density (HPD) interval for the mutation and recombination treatment effects. We evaluate the coverage pointwise for each nucleotide position of the 600 bp long sequences. Although the nominal level of 95% is not achieved in either experiment, the observed coverage rates are encouraging given the small size of the sequence datasets.

Figure 4 shows the coverage of the 95% HPD intervals of $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ (top left) and $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$ (top right) in the first experiment. The coverage of the true mutation treatment effect (top left) is lowest in the region with the larger mutation rate, suggesting the model is conservative in these elevated mutation rates in the presence of the surrounding lower mutation rates. The coverage of the true recombination treatment effect (top right) is high except at the boundaries of the region with elevated mutation treatment effect. The breakpoints of this central region are the most difficult areas for the model to estimate accurately since the increased sequence changes in this central region can be attributed to either the recombination or mutation processes. The overall coverage across all positions in experiment 1 was 87% for the mutation treatment effect and 89% for the recombination treatment effect.

Figure 4 also shows the coverage of the 95% HPD intervals of $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ (bottom left) and $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$ (bottom right) in the second experiment which has true nonzero mutation and recombination treatment effects. Similar to Experiment 1, we observe that the coverage of the true mutation treatment effect (bottom left) is lowest in the region with the elevated mutation rate. However, the interval coverage for mutation and recombination rates is not degraded further by the presence of a nonzero recombination treatment effect in Experiment 2. The overall coverage across all positions in Experiment 2 was 90% for the mutation treatment effect and 87% for the recombination treatment effect.

3.2 Simulation Evaluation: Underlying Evolution Parameters

With the next set of simulation studies, we evaluate the estimation of our global parameters $(\mu_0, \rho_0, \sigma_\mu^2, \sigma_\rho^2, \tau_\mu^2, \tau_\rho^2, \kappa)$ and within-block mutation and recombination parameters $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$. In this set of simulations, we used a fixed blocking structure so that it is easier to evaluate our ability to capture the variance between blocks σ_μ^2 and σ_ρ^2 as well as the within-block rates $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$.

We consider a variety of different data settings in this simulation study, which we enumerate in Table 1. The first three settings (A–C) explore different levels of mutation and recombination with a limited number of short sequences. In settings D and E, we expand the length of each sequence in our dataset as well as the number of blocks of different mutation and recombination rates. Finally, in setting F, we expand both the length of each sequence as well as the number of sequences in both treatment and control datasets. In terms of the number of sequences and length of sequences, Setting F is similar to the datasets we see in our applications in Section 4.

For all conditions in Table 1, we generated 20 treatment and control sequence datasets and evaluated the coverage of

Table 1. Description of simulation conditions for studying our underlying evolution parameters

Data setting	Sequence length	Number of blocks	Number of seqs	Mutation levels	Recombination levels
A	160 bp	4	5	High	Low
B	160 bp	4	5	Low	High
C	160 bp	4	5	High	High
D	1000 bp	10	5	High	Low
E	1000 bp	10	5	Low	High
F	2000 bp	10	20	High	Low

NOTE: The "high" setting is a rate of $\mu_0 = 0.05$ or $\rho_0 = 0.05$ and the "low" setting is a rate of $\mu_0 = 0.005$ or $\rho_0 = 0.005$. For "high" rates, variances were set to $\sigma^2 = 0.25$ and $\tau^2 = 0.05$. For "low" rates, variances were set to $\sigma^2 = 0.05$ and $\tau^2 = 0.001$. In all conditions, κ was set to 0.5.

the 95% posterior intervals for each parameter in our model. In Table 2, we give the coverage for each global parameter $(\mu_0, \rho_0, \sigma_\mu^2, \sigma_\rho^2, \tau_\mu^2, \tau_\rho^2, \kappa)$ as well as the average coverage of each within-block central rate μ_i and ρ_i across all blocks i. In our supplementary materials, we also give results for the root mean square errors (RMSEs) of the posterior means inferred from our model.

Overall, we see in Table 2 that the global mutation and recombination rates parameters (μ_0 and ρ_0) have better coverage compared to the within-block mutation and recombination parameters (μ_i and ρ_i). This result was expected since there is less information in the data for the parameters in each individual block compared to the global parameters. In our supplementary materials, we also see higher RMSEs for the within-block mutation and recombination parameters compared to the global rates μ_0 or ρ_0.

The coverage rates from our model are poor for the within-block mutation and recombination parameters (μ_i and ρ_i) in the small data settings A, B, and C which only have 40 nucleotides per block. The coverage rates of the within-block mutation and recombination parameters improve when we move to the intermediate data settings, D and E, which have 100 nucleotides per block. The coverage rates are quite good for the within-block mutation and recombination parameters (μ_i and ρ_i) in the larger data setting F that most closely matches the applications of Section 4.

However, even in setting F, we note that the variance parameters (especially σ_ρ^2 and τ_ρ^2) do not have good coverage, which suggests that detailed information about the recombination process is difficult to capture even with larger datasets. We will also see in Section 4 that recombination is difficult to estimate in real data. In our supplementary materials, we show that the RMSEs for the variance parameters are also poor in all settings (including the larger data setting F).

Table 2. Coverage rates of true parameter values by 95% posterior intervals from our model across 20 datasets generated under the settings outlined in Table 1

Setting	μ_0	ρ_0	σ_μ^2	σ_ρ^2	τ_μ^2	τ_ρ^2	κ	μ_i	ρ_i
A	0.95	0.90	1.00	0.95	0.25	0.55	0.75	0.30	0.56
B	1.00	0.75	1.00	0.95	0.70	0.45	0.85	0.59	0.44
C	0.85	0.70	0.90	0.95	0.25	0.25	0.65	0.43	0.38
D	0.90	0.90	0.90	0.90	0.60	0.70	0.50	0.70	0.58
E	1.00	0.85	0.95	1.00	0.60	0.30	0.80	0.62	0.48
F	1.00	1.00	1.00	0.40	0.85	0.20	1.00	0.97	1.00

NOTE: Note that the coverage rates in the final two columns are also averaged across all blocks.

Table 3. False-positive and false-negative rates for mutation $(\mu_i^T - \mu_i^C)$ and recombination $(\rho_i^T - \rho_i^C)$ treatment effects

False-negative rates

Data size	Effect size	Mutation	Recombination
Small	Small	0.91	0.98
Small	Large	0.78	0.93
Large	Small	0.94	0.67
Large	Large	0.74	0.75

False-positive rates

Data size	Mutation	Recombination
Small	0.04	0.04
Large	0.50	0.23

NOTE: "Large" data size consists of simulation setting F in Table 2 whereas "small" data size consists of simulation settings A–E in Table 2. We used the upper third and lower third of our true treatment effects to define "large" versus "small" treatment effects.

We also evaluated our ability to estimate treatment effects within the simulation settings outlined in Table 2. We first define the "detection" of a treatment effect as a 95% posterior interval for that treatment effect which does not contain zero and has the same sign as the true treatment effect. Across our simulated datasets, we tabulate 1. false-positive (FP) events where our model detects a nonzero treatment effect when the true treatment effect is actually zero, and 2. false-negative (FN) events where our model fails to detect a treatment effect when the true treatment effect is nonzero. In Table 3, we give the FP rate and FN rates of our model for both mutation and recombination treatment effects. We calculated FP and FN rates separately for our small data simulations (settings A–E) versus large data simulations (setting F). False-negative rates were also calculated separately for small values versus large values of the true treatment effects.

Table 3 shows different trends for the false-negative rates for mutation treatment effects versus recombination treatment effects. For the mutation treatment effects, the size of the effect makes the biggest difference for the FN rate. Larger effect sizes lead to lower false-negative rates, whereas the size of the dataset (large vs. small) makes less of a difference. In contrast, the false-negative rates for recombination treatment effects are substantially reduced in large datasets versus small datasets, but the size of the treatment effect itself makes less of a difference for recombination. In our supplementary materials, we give further details of the false-negative rates (i.e., power) of our procedure as a function of effect size for both mutation and recombination treatment effects.

The false-positive rate is very low in our small data setting which can be attributed to the conservative nature of our model, which shrinks treatment effects toward zero when there is not a strong data signal. When larger amounts of data are available, our model is better able to detect true differential treatment

effects, but there is a corresponding increase in the FP rate in the larger data settings.

In summary, our simulation results suggest that the performance of our model, in terms of accurately capturing the underlying evolutionary parameters, is quite sensitive to the amount of the available sequence data. Our model is conservative in smaller data settings and it is reassuring to see good coverage of most parameters (including within-block mutation and recombination rates) in settings that more closely match the size of our real data applications. The detection of treatment effects (differential mutation or recombination) is sensitive to both the size of the data as well as the size of the true treatment effect. Our procedure behaves conservatively in small data settings by shrinking treatment effects toward zero. In our supplementary materials, we give further details about the coverage and detection of treatment effects.

4. APPLICATIONS: DRUG AND GENE THERAPIES

We apply our model to estimating the evolutionary escape response for two different HIV therapies: one involving the drug Enfuvirtide, and the other a VIRxSYS antisense gene therapy. For the latter application, we also compare our results with simpler approaches. In all of these studies, we focus on treatment effects $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{C}$ and $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$. These effects are the spatially varying differences in recombination rates and mutation rates inferred by our model, which we interpret as the evolutionary escape response of HIV.

4.1 Application: Enfuvirtide Drug Therapy

In our first application, we examine a therapy based on the drug Enfuvirtide, also known as Fuzeon or T-20 (Wild, Greenwell, and Matthews 1993). This drug is designed to attack the protein coat of the HIV virions outside of human cells, preventing the virus from binding to the cell surface and delivering its viral load into the host cell. Even a small number of HIV mutations can lead to drug resistance by producing subtle changes in the HIV protein coat that prevent Enfuvirtide targeting. Recombination serves as a mechanism for virions in the HIV population to share beneficial mutations.

In this study, blood samples were taken before exposure to drug (control) and after exposure to drug (treatment). The existence of a significant postexposure HIV population in these patients indicates the evolution of sequence changes that conferred resistance to the action of the drug therapy. Aligned sequences of the *env* (envelope protein) region of the HIV genome were generated from these blood samples.

Analyzing each patient individually was not a viable option since a very small number of sequences (5–8) were available from each patient. Instead, we pooled sequences across all five patients, giving us 28 control sequences and 29 treatment sequences. Although necessary, this pooling violates our model assumption that the set of sequences in the treatment group (or control group) share a common genealogy, which might impair our ability to estimate differential mutation or recombination between the treatment and control groups.

In Figure 5, we plot the posterior mean (blue line), posterior median (purple line), and the 95% HPD intervals (black lines) for the mutation treatment effect $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$. The left plot in Figure 5 shows the entire length of the *env* region in our dataset, where substantial spatial heterogeneity in the mutation treatment effect is observed.

The right plot in Figure 5 magnifies the region of nucleotide positions 1590–1700, the only region where the entire 95% HPD interval for $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ is above zero. As indicated in the plot, this region of significantly elevated mutation rates in the treatment population is also known to be the location of resistance-conferring mutations (Ray et al. 2007).

In light of false-positive rates seen in our simulation evaluation, we are cautious about attributing the elevated mutation rates directly to the action of the drug therapy. However, the overlap with known resistance-conferring mutations does at least

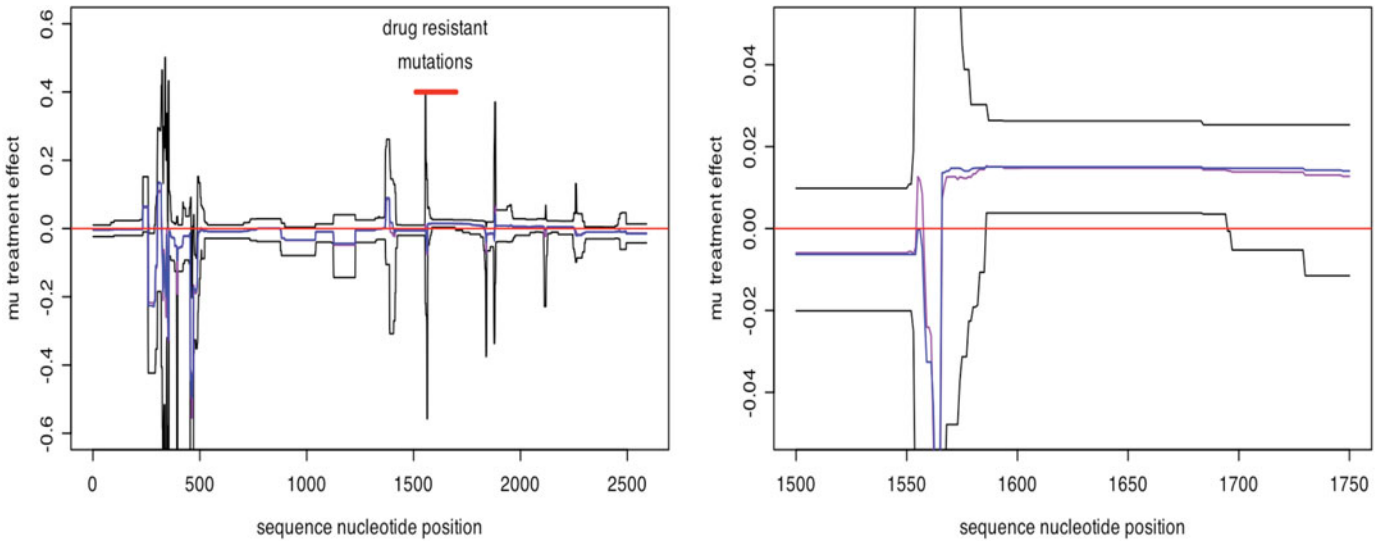

Figure 5. Left: mutation treatment effect $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ of enfuviritide drug therapy for the entire *env* sequence. Right: subsection of *env* sequence with significant mutation treatment effect. Blue line is posterior mean, purple line is the posterior median, and black lines are 95% highest-posterior-density (HPD) intervals. The online version of this figure is in color.

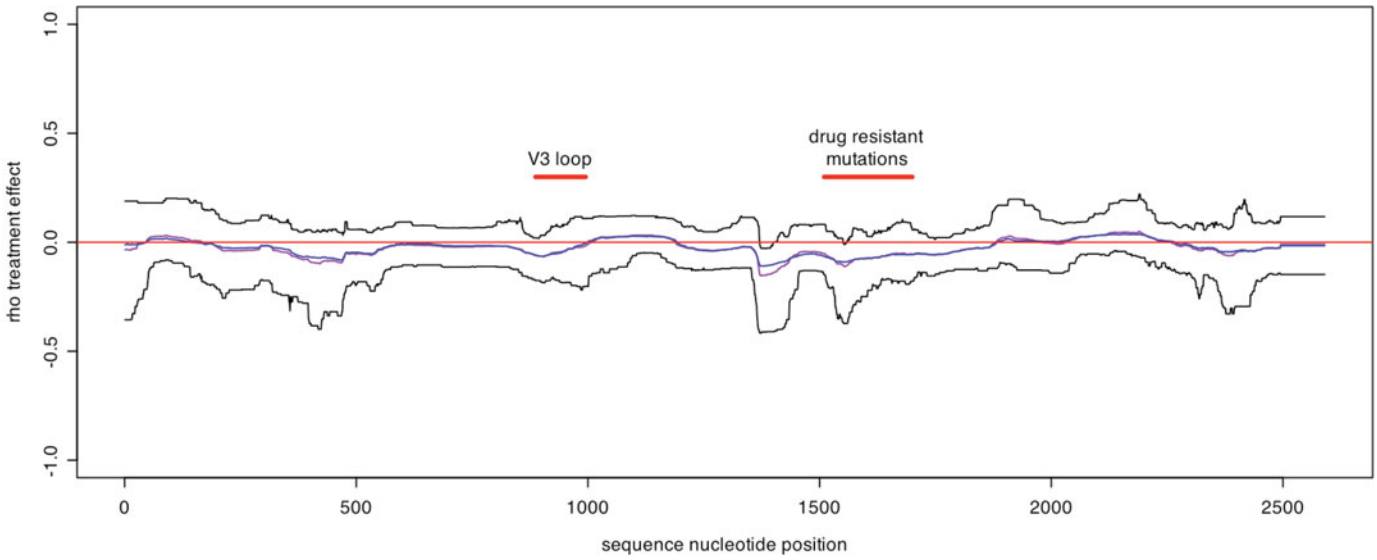

Figure 6. Recombination treatment effect $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$ of enfuviritide drug therapy for the entire *env* sequence. Blue line is posterior mean, purple line is the posterior median, and black lines are 95% highest-posterior-density (HPD) intervals. The online version of this figure is in color.

suggest that our method is capturing biologically relevant signal. It is also notable that a region of significantly elevated mutation was inferred despite the conservative shrinkage of our model observed in Section 3.2 for smaller data settings.

We also observe increased posterior variability in the region of nucleotide positions 300–500, which may be attributed to the fact that this region overlaps with the location of the V1/V2 loops of the *env* sequence. The V1/V2 loops are a highly variable region of the HIV genome lying between amino acids 119 and 205, which corresponds to nucleotides 357–615 in the *env* sequence (Wyatt et al. 1995).

In Figure 6, we plot the posterior mean (blue line), posterior median (purple line), and the 95% HPD intervals (black lines) for the recombination treatment effect $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$. We have external knowledge that both the treatment and control populations were mixtures of different HIV subtypes, and that the same resistance-conferring mutations occurred in HIV sequences with different subtypes (Ray et al. 2007). It is possible that these resistance-conferring mutations were passed between the different subtypes via recombination.

Subtype identity is specified in the V3 loop region of the *env* sequence, which corresponds to nucleotide positions 887–995. We expect that some amount of recombination must be occurring in the region between the V3 loop region and the resistance-conferring mutations, both of which are labeled in Figure 6. We see some spatial heterogeneity in the differential recombination rate between treatment and control in Figure 6 but no regions of significantly elevated recombination.

Since recombination can occur at any point in the 600 bp sequence between the V3 loop region and the resistance-conferring mutations, it might not be surprising that our model does not detect a concentrated area of elevated recombination, especially given the conservative shrinkage of our model observed in Section 3.2 in smaller data settings. Our ability to detect elevated recombination rates could also be compromised by the model assumption of a common genealogy, given that sequences were pooled across multiple patients.

4.2 Application: VIRxSYS Antisense Gene Therapy

In this study, an HIV population was exposed in vitro to an antisense gene therapy developed by the company VIRxSYS. This new type of therapy differs substantially from drug therapies that attack HIV virions outside of human cells, acting instead after the HIV virion has infected a host cell. Within the cytosol of the host, the antisense therapy binds directly to the HIV genome, creating a double-stranded RNA which is recognized as foreign and degraded by normal cellular functions. The different mechanisms of drug and antisense therapies are contrasted in Figure 7.

A promising feature of antisense gene therapy is that a larger region of the HIV genome is directly targeted compared to drug therapies, so substantially more sequence changes should be needed to develop resistance. This particular VIRxSYS

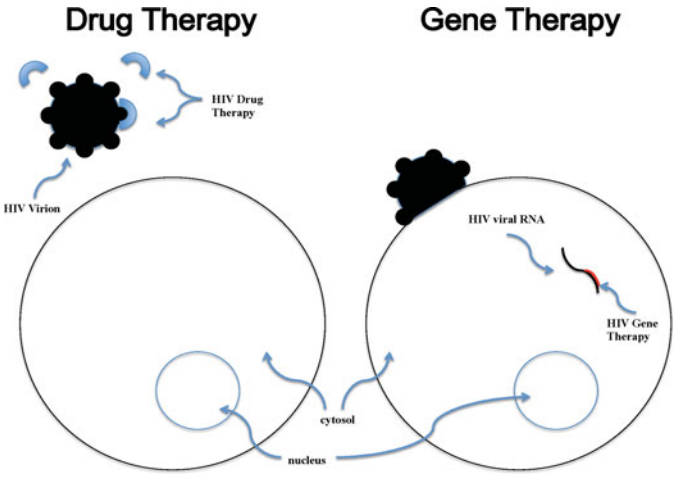

Figure 7. Left: drug therapies attack the HIV virion prior to infection of human cell. Right: antisense gene therapies attack the HIV viral genome after infection of human cell. The online version of this figure is in color.

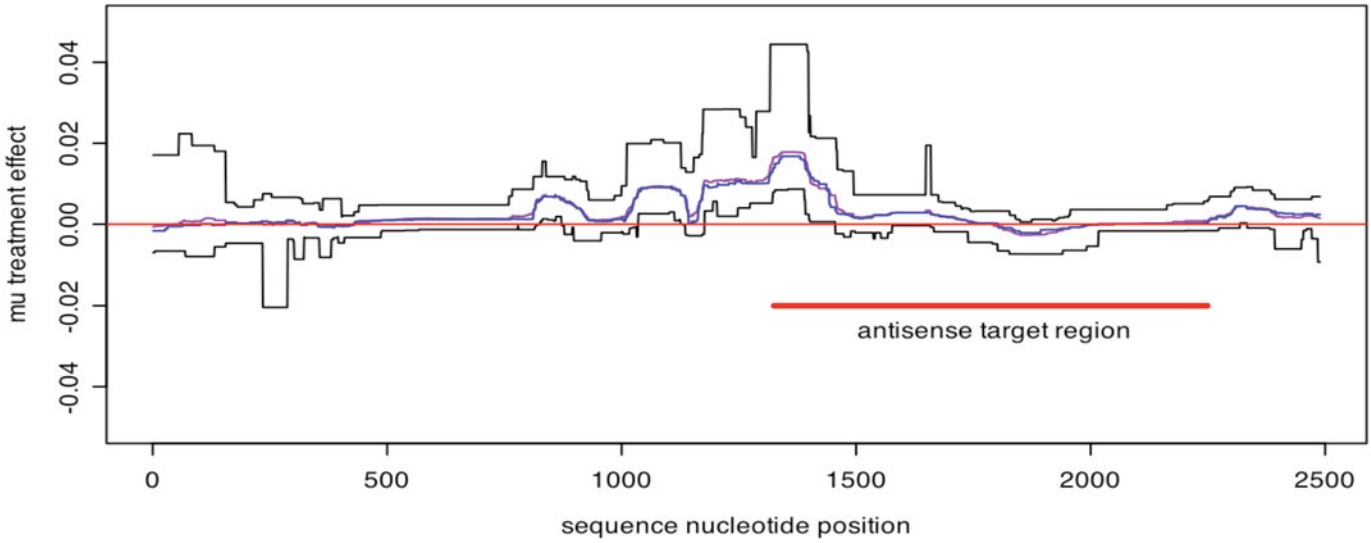

Figure 8. Mutation treatment effect $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ of the VIRxSYS antisense gene therapy on mutation rates for the entire *env* sequence. Blue line is posterior mean, purple line is the posterior median, and black lines are 95% highest-posterior-density (HPD) intervals. The online version of this figure is in color.

antisense therapy targets a large region (nucleotides 1325–2249) of the HIV *env* gene (Lu et al. 2004).

The dataset consists of 48 aligned treatment sequences from an HIV population in cell culture that was exposed to the antisense vector and 19 aligned control sequences from an HIV population in a separate neutrally evolving cell culture that was not exposed to therapy. In this setting, it is reasonable to assume common genealogies for the sequences within the treatment and control groups that are also independent between the two groups.

In Figure 8, we examine the posterior mean (blue line), posterior median (purple line), and the 95% HPD intervals (black lines) for the mutation treatment effect $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$.

Similar to the drug therapy, there is substantial overall spatial heterogeneity in the mutation treatment effect for the antisense gene therapy. The main feature of Figure 8 is the large contiguous region of significantly elevated mutation in the center of the *env* sequence. The elevated mutation rates are highest at the 5′ boundary of the antisense target region (labeled in Figure 8) and overlap the target region for approximately 425 nucleotides in the 3′ direction.

Figure 9 gives the posterior mean (blue line), posterior median (purple line), and the 95% HPD intervals (black lines) for the recombination treatment effect $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$ of the VIRxSYS antisense gene therapy. We do not see a substantial recombination treatment effect.

Although sequence changes in the antisense target region were expected, it was not clear a priori whether mutation, recombination, or both would be the dominant mechanism of the evolutionary escape response. A contribution of our analysis is the putative finding that mutation rates are significantly elevated in the treatment group, suggesting that mutation is a primary factor in HIV's response to the VIRxSYS antisense gene therapy.

It is unsurprising that a much larger region of elevated mutation is needed to escape the antisense gene therapy (Figure 8)

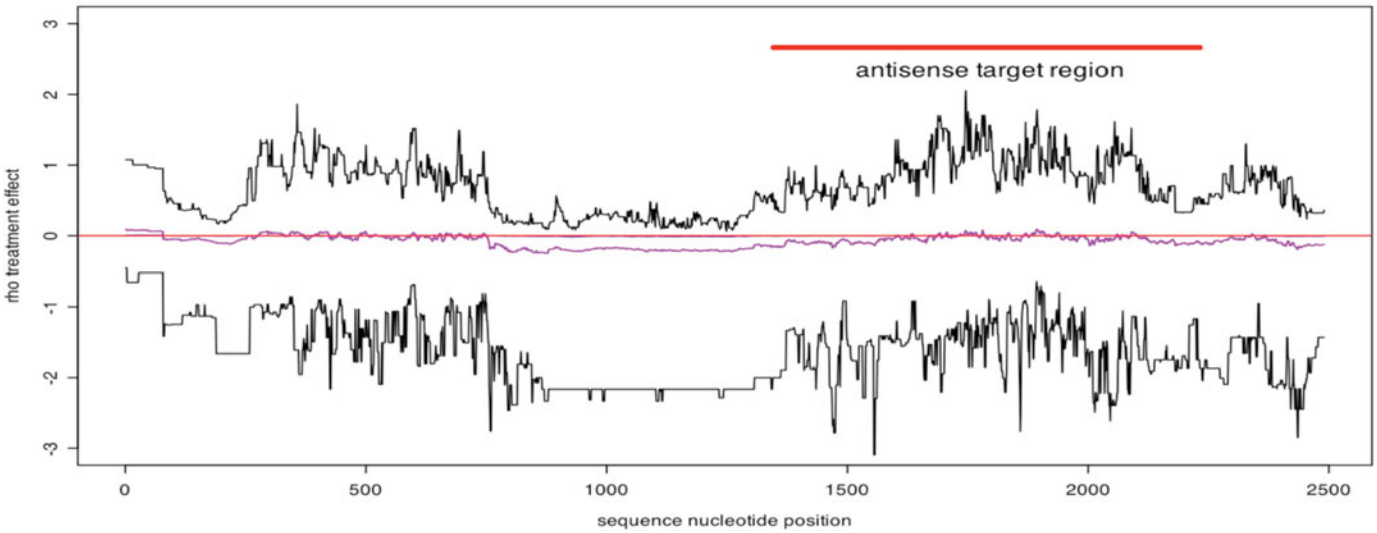

Figure 9. Treatment effect $\boldsymbol{\rho}^{\mathrm{T}} - \boldsymbol{\rho}^{\mathrm{C}}$ of VIRxSYS antisense gene therapy on recombination rates for the entire *env* sequence. Blue line is posterior mean, purple line is the posterior median, and black lines are 95% highest-posterior-density (HPD) intervals. The online version of this figure is in color.

compared to the drug therapy (Figure 5), since that drug's effect can be inhibited by a smaller number of well-placed mutations.

However, it was not expected a priori that the region of elevated mutation in Figure 8 would be shifted relative to the antisense target region. Although we must exercise caution about the possibility of a false discovery, this shift in the antisense target region is a potentially novel finding. We speculate that the shift could be a consequence of adenosine deaminases that act on RNA (ADARs). ADARs induce A to G changes in double-stranded RNA (Bass 2002; Nishikura 2010), which would be produced any time the antisense therapy binds to the HIV genome.

ADARs have been previously implicated in the response to antisense gene therapy (Lu et al. 2004; Mukherjee et al. 2010). However, although there have been recent analyses of ADAR editing locations (Wulff, Sakurai, and Nishikura 2011), the actual mechanism of the ADAR editing is poorly understood. We hope that further work in this area will confirm our putative results and elucidate the role of ADAR in the HIV response to antisense gene therapies.

4.3 Simpler Analyses of VIRxSYS Antisense Gene Therapy

In this section, we examine simpler alternatives to our model for analyzing the evolutionary escape response to the VIRxSYS antisense gene therapy. A natural starting point is the examination of *segregating sites*, which are the nucleotide positions in the genome where at least one sequence has a different nucleotide than the other sequences. A simple analysis would be a comparison of the segregating sites in the treatment HIV sequences versus the control HIV sequences.

In Figure 10, we plot the spatial position of all segregating sites in the treatment sequences (blue) and control sequences (red). For comparison, we also include on top the posterior distribution of the mutation treatment effect $\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{C}}$ estimated by our model (from Figure 8).

The treatment sequences contain a higher frequency of segregating sites in the region consisting of positions 800–1500 in Figure 10. This same region corresponds to the elevated mutation treatment effect estimated by our model. However, the examination of segregating sites cannot distinguish between mutation and recombination processes as we do with our hierarchical model in Section 4.2.

We also explored using segregating-site information as input to a prediction model for the differences between the treatment and control HIV populations. The outcome variable for this model is the treatment or control label of each sequence. Segregating sites that are involved in the evolutionary escape response of HIV should also be associated with the treatment labeling. There are 189 indicator covariates for each individual segregating site as well as 51 "pooled" covariates that tabulate the number of segregating sites in blocks of 50 contiguous nucleotides.

We used a Bayesian additive regression tree (BART) model (Chipman, George, and McCulloch 2008), which is well suited for prediction when the number of covariates exceeds the number of observed sequences. BART is a Bayesian tree ensemble model, where each tree in the ensemble is constrained by a prior distribution to be a weak learner. The covariates that appear most often in the ensemble tend to show the strongest association with differences between treatment and control populations. Of the segregating sites that were frequently included in the BART ensemble, the majority were located in the region consisting of nucleotide positions 800–1700, which roughly corresponds to the elevated mutation region found by our model in Figure 8.

These simpler segregating-site analyses lack several important features of our model-based approach (Section 2). Our model estimates differences between treatment and control

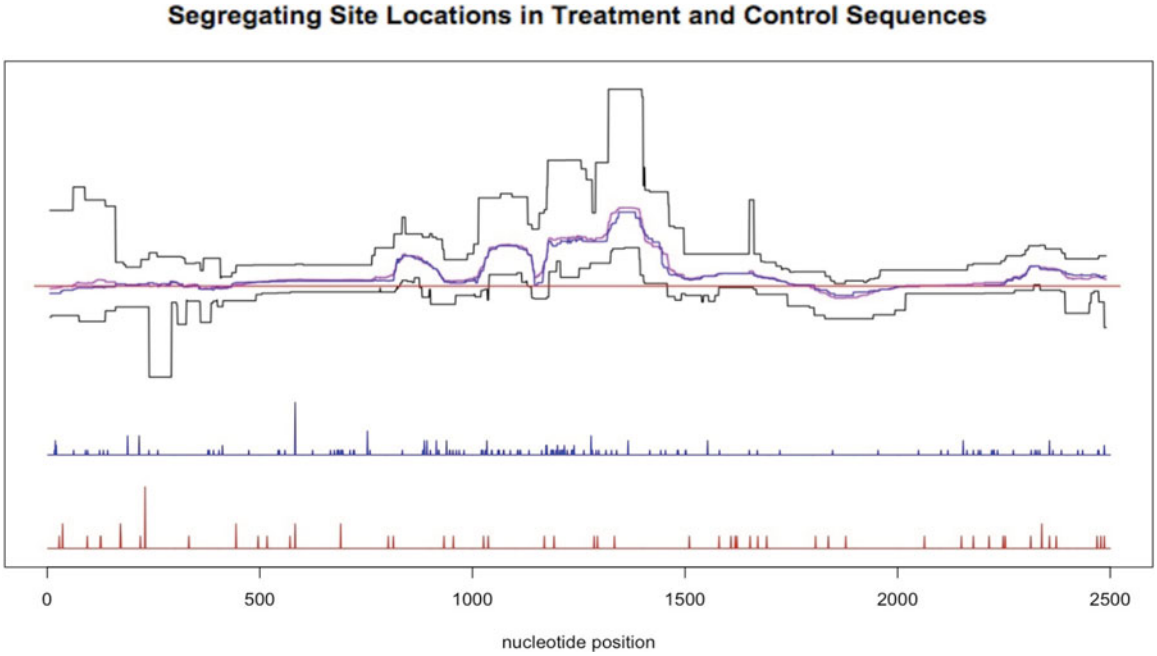

Figure 10. Spatial positions of segregating sites in treatment sequences (blue spikes) and in control sequences (red spikes). Height of spike is proportional to the proportion of sequences that do not have the most common nucleotide. For comparison, the posterior VIRxSYS mutation treatment effect from Figure 8 is also overlaid on the plot. The online version of this figure is in color.

populations in terms of both mutation and recombination processes, whereas segregating sites do not distinguish between these two mechanisms of sequence change. Our hierarchical Bayesian approach shares information between the groups as well as proximal regions of the HIV genome, while providing posterior variances for estimated treatment effects.

5. DISCUSSION

We have introduced a Bayesian hierarchical model for the estimation of evolutionary escape response in an HIV population exposed to medical therapy. The evolution of the population is modeled using mutation and recombination rate parameters that differ between treatment and control populations, while still sharing information through a hierarchical framework. Both mutation and recombination processes must be modeled simultaneously, as both processes for sequence change occur frequently in HIV.

Our model allows for spatial heterogeneity in the evolutionary response, through a piecewise-constant blocking prior on the mutation and recombination parameter profiles. This is a crucial aspect of our approach, since different regions of the HIV genome are generally under different evolutionary pressure due to the targeting of each therapy.

We applied our model to two HIV treatments: a drug therapy and an antisense gene therapy. Although the goal of both therapies is the disruption of the HIV infection cycle, they have fundamentally different mechanisms of action. In both of these applications, we found *putative* evidence of HIV evolution in response to therapeutic pressure. We emphasize *putative* since our simulation evaluation does suggest our method can produce false positives in data settings of similar size to our real data applications.

For the enfuvirtide drug therapy, our model detected elevated mutation rates in a small region that corresponds to the location of known drug-resistant mutations, and we also see elevated posterior variance in the V1/V2 variable loop region. For the VIRxSYS antisense gene therapy, our model detected a much larger region of significantly elevated mutation rates which overlaps with the target region of the antisense vector. The fact that the elevated mutation region is shifted relative to the target region of the antisense vector is a potentially important result.

Although motivated by HIV therapies, our methodology could be useful for any two-group comparison of sequence data where evolutionary differences are suspected to have occurred. It is especially relevant for populations exhibiting significant recombination, since simpler models which ignore recombination can produce misleading results (Schierup and Hein 2000).

Extending our methodology to more complicated mutation models is a subject of ongoing research. One weakness of our approach is that the MCMC implementation is quite computationally intensive. Alternative model estimation strategies, such as variational inference, could provide fruitful future directions for methodology in this area.

SUPPLEMENTARY MATERIALS

Root mean square errors of parameter estimates, coverage of treatment effects, and power for detecting treatment effects from our simulation study in Section 3.2 are posted online as supplementary materials.

[Received April 2011. Revised July 2013.]

REFERENCES

Bass, B. L. (2002), "RNA Editing by Adenosine Deaminases That Act on RNA," *Annual Review of Biochemistry*, 71, 817–846. [1241]

Chipman, H., George, E., and McCulloch, R. (2008), "BART: Bayesian Additive Regression Trees," *Annals of Applied Statistics*, 4, 266–298. [1241]

Fearnhead, P., and Donnelly, P. (2001), "Estimating Recombination Rates From Population Genetic Data," *Genetics*, 159, 1299–1318. [1231]

Felsenstein, J., and Churchill, G. (1996), "A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution," *Molecular Biology and Evolution*, 13, 93–104. [1231]

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721–741. [1233]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–731. [1233]

Griffiths, R., and Marjoram, P. (1996), "Ancestral Inference From Samples of DNA Sequences With Recombination," *Journal of Computational Biology*, 3, 479–502. [1230,1232]

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109. [1233]

Hein, J., Schierup, M. H., and Wiuf, C. (2005), *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*, New York: Oxford University Press. [1230]

Hudson, R. (1983), "Properties of a Neutral Allele Model With Intragenic Recombination," *Theoretical Population Biology*, 23, 183–201. [1230]

——— (2002), "Generating Samples Under a Wright-Fisher Neutral Model," *Bioinformatics*, 18, 337–338. [1235]

Kingman, J. (1982), "The Coalescent," *Stochastic Processes and Their Applications*, 13, 235–248. [1230,1231]

Li, N., and Stephens, M. (2003), "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data," *Genetics*, 165, 2213–2233. [1231,1232]

Lu, X., Yu, Q., Binder, G. K., Chen, Z., Slepushkina, T., Rossi, J., and Dropulic, B. (2004), "Antisense-Mediated Inhibition of Human Immunodeficiency Virus (HIV) Replication by Use of an HIV Type 1-Based Vector Results in Severely Attenuated Mutants Incapable of Developing Resistance," *Journal of Virology*, 78, 7079–7088. [1240,1241]

McGuire, G., Denham, M., and Balding, D. (2001), "Models of Sequence Evolution for DNA Sequences Containing Gaps," *Molecular Biology and Evolution*, 18, 481–490. [1231]

Mukherjee, R., Plesa, G., Sherrill-Mix, S., Richardson, M. W., Riley, J. L., and Bushman, F. D. (2010), "HIV Sequence Variation Associated With *env* Antisense Adoptive T-Cell Therapy in the hNSG Mouse Model," *Molecular Therapy*, 18, 803–811. [1241]

Nishikura, K. (2010), "Functions and Regulation of RNA Editing by ADAR Deaminases," *Annual Review of Biochemistry*, 79, 321–349. [1241]

Rabiner, L. R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–286. [1232]

Rambaut, A., and Grassly, N. C. (1997), "Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution Along Phylogenetic Trees," *Computer Applications in the Bioscience*, 13, 235–238. [1235]

Ray, N., Harrison, J. E., Blackburn, L. A., Martin, J. N., Deeks, S. G., and Doms, R. W. (2007), "Clinical Resistance to Enfuvirtide Does Not Affect Susceptibility of Human Immunodeficiency Virus Type 1 to Other Classes of Entry Inhibitors," *Journal of Virology*, 81, 3240–3250. [1238,1239]

Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H. (1995), "Recombinaton in HIV-1," *Nature*, 374, 124–126. [1230]

Schierup, M. H., and Hein, J. (2000), "Consequences of Recombination on Traditional Phylogenetic Analysis," *Genetics*, 156, 879–891. [1230,1242]

UNAIDS (2006), *Aids Epidemic Update: Special Report on HIV/AIDS, Joint United Nations Programme on HIV/AIDS*, Geneva, Switzerland: UNAIDS. [1230]

Wild, C., Greenwell, T., and Matthews, T. (1993), "A Synthetic Peptide From HIV-1 gp41 is a Potent Inhibitor of Virus Mediated Cell-Cell Fusion," *AIDS Research and Human Retroviruses*, 9, 1051–1053. [1238]

Wilson, D. J., and McVean, G. (2006), "Estimating Diversifying Selection and Functional Constraint in the Presence of Recombination," *Genetics*, 172, 1411–1425. [1231,1232,1233]

Wulff, B.-E., Sakurai, M., and Nishikura, K. (2011), "Elucidating the Inosinome: Global Approaches to Adenosine-to-Inosine RNA Editing," *Nature Reviews*, 12, 81–85. [1241]

Wyatt, R., Moore, J., Accola, M., Desjardin, E., Robinson, J., and Sodroski, J. (1995), "Involvement of the v1/v2 Variable Loop Structure in the Exposure of Human Immunodeficiency Virus Type 1 gp120 Epitopes Induced by Receptor Binding," *Journal of Virology*, 69, 5723–5733. [1239]