A Bayesian Framework for Inferring the Influence of Sequence Context on Point Mutations

Guy Ling,[†,1] Danielle Miller,[†,1] Rasmus Nielsen,[2,3,4] and Adi Stern[*,1,5]

[1]School of Molecular Cell Biology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel

[2]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA

[3]Department of Statistics, University of California, Berkeley, Berkeley, CA

[4]Center for Computational Biology at UC Berkeley (CCB), Berkeley, CA

[5]Edmond J. Safra Center for Bioinformatics at Tel Aviv University, Tel-Aviv, Israel

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: sternadi@tau.ac.il.

Associate editor: Claus Wilke

Abstract

The probability of point mutations is expected to be highly influenced by the flanking nucleotides that surround them, known as the sequence context. This phenomenon may be mainly attributed to the enzyme that modifies or mutates the genetic material, because most enzymes tend to have specific sequence contexts that dictate their activity. Here, we develop a statistical model that allows for the detection and evaluation of the effects of different sequence contexts on mutation rates from deep population sequencing data. This task is computationally challenging, as the complexity of the model increases exponentially as the context size increases. We established our novel Bayesian method based on sparse model selection methods, with the leading assumption that the number of actual sequence contexts that directly influence mutation rates is minuscule compared with the number of possible sequence contexts. We show that our method is highly accurate on simulated data using pentanucleotide contexts, even when accounting for noisy data. We next analyze empirical population sequencing data from polioviruses and HIV-1 and detect a significant enrichment in sequence contexts associated with deamination by the cellular deaminases ADAR 1/2 and APOBEC3G, respectively. In the current era, where next-generation sequencing data are highly abundant, our approach can be used on any population sequencing data to reveal context-dependent base alterations and may assist in the discovery of novel mutable sites or editing sites.

Key words: **mutation rates, sequence context, population genetics, evolutionary model, MCMC.**

Introduction

Single-base modifications, which include point mutations and DNA or RNA modifications, are often caused by enzymatic activity. Base alterations can include either standard point mutations or modifications such as cytosine methylation at the DNA level (Cooper and Krawczak 1989), or adenine to inosine at the RNA level (Wulff et al. 2011). For DNA/RNA modifications, specific sequence contexts influence the probability the enzyme will modify a base within this context (Lehmann and Bass 2000; Feltus et al. 2006; Wang et al. 2013). This is reflected by hotspots of mutation in different genomes driven by specific sequence contexts (Hwang and Green 2004; Hodgkinson and Eyre-Walker 2011; Aggarwala and Voight 2016). One well-known example is that of C→T mutations that occur at high rates in vertebrate genomes by spontaneous deamination of methylated cytosines in CpG positions, that is, positions in which a guanine follows a cytosine (Coulondre et al. 1978; Razin and Riggs 1980). Moreover, specific cellular enzymes belonging to the APOBEC3 family increase the rate of deamination of cytosine bases as a means of viral restriction. These enzymes increase the mutation rate of HIV by several orders of magnitude, at specific sequence contexts (Cuevas et al. 2015). Nevertheless, most evolutionary models commonly assume that every position evolves independently. This implies that neighboring positions do not affect the rate of mutation in a given position. Here we present a novel Bayesian method for the analysis of deep population sequencing data, which detects the effect of context on the rate of single-base modifications.

Previous efforts for detecting the effects of context on substitution rates were often phylogenetic-based methods (Siepel and Haussler 2003). In order to keep the number of parameters small and computationally tractable, most such methods consider two to three sites (Krawczak et al. 1998; Dunson and Tindall 2000; Lunter and Hein 2004; Hernandez et al. 2007; Zhang et al. 2007; Rodrigue et al. 2009, 2010; Simmonds et al. 2013; Sung et al. 2015; Figliuzzi et al. 2016; Harris and Pritchard 2017) (but see Berikov and Rogozin 1999). Notably, methods which attempt to find both local and global explicit context dependencies usually suffer from

Article

high false positive rate, and hence require more statistical validations (Rogozin et al. 2005).

Recently, novel models for analyzing context from polymorphism data in humans were suggested (Aggarwala and Voight 2016; Zhu et al. 2017). However, these models are very high-dimensional, and in essence, assume that a very large number of sequence motifs may influence the rate of substitution. An additional complication is that within the high-dimensional space of sequence motifs, there is a high level of correlation between motifs, which could lead to faulty inferences of context effects (but see Aikens et al. 2019). For example, the motif **CGC**XX is contained within the motif X**GC**XX (where X is any one of the four nucleotides) and this could lead to difficulty in inferring which motif truly has an effect.

Here, we develop a model that addresses these problems and includes two main novelties. First, it is tailored for next-generation sequencing (NGS) of population data, which are becoming more and more abundant. Notably, our method is inspired by ours and other experiments that sequence virus populations at great depth (Acevedo et al. 2014; Stern et al. 2017); typically such experiments result in over 100,000–1,000,000 sequenced viral genomes, with sequencing accuracy that allowed detection of mutations present at a frequency as rare as 10^{-6}. Moreover, in these experiments the very high mutation rate of the viruses is the predominant force: Genetic drift is mitigated by the large population size (but see Discussion), and the short time frame of the experiment mostly allows for only one mutation per genome. The second novelty is the use of Bayesian variable selection methods to identify the few sequence contexts that significantly influence the substitution rate, thereby addressing the combinatorial increase in parameter number with a larger context. This approach relies on the biological motivation that only a limited number of enzymes influence the rate of base alteration and they are often defined by a very specific context. For example, adenosines in mRNA may be methylated by a methyltransferase enzyme but only in the context RGACU (where R is a purine) (Narayan et al. 1994). There are over 1,000 possible contexts in a window size of 5 ($k = 5$), but as only two of them exert an effect, this is an assumption worthwhile taking into consideration.

The problem of detecting contexts that affect the rate of base alteration is a special case of the standard statistical problem of identifying a subset of, possibly correlated, covariates that affect the response variable. One example is quantitative trait locus (QTL) mapping, where the objective is to detect a limited set of markers that affect a specific phenotype, and our method is inspired by Bayesian variable selection solutions to this problem (see, e.g., Yi 2004). We test the method on simulated data generated to mimic an increase or decrease of mutation rate caused by specific nucleotide motifs, in a population of replicating viruses, and show that the method accurately captures these changes in the simulated data. We next analyze data from NGS experiments of polioviruses and HIV-1 and describe how our method captures intriguing biological signals. Finally, we discuss the applicability of our method to any type of NGS data.

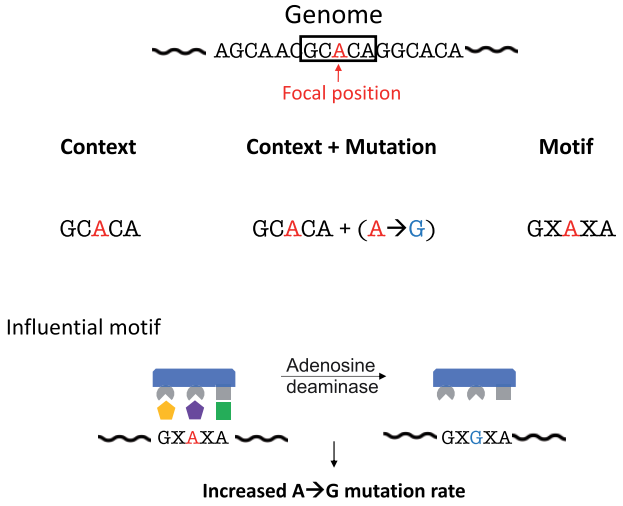

FIG. 1. Model definitions. For the labeled focal position with nucleotide A, we present its genomic context for $k = 5$, and illustrate the context, the context together with a mutation, and a possible motif embedded in this context. Motifs are associated with a context pattern, such that $X \in \{A, C, G, T/U\}$. We exemplify how enzymatic activity operating on a specific sequence context may result in an increase in the mutation rate at this context.

Theory

In brief, our method searches for a correlation between the presence of sequences motifs surrounding a site and elevated/decreased mutation frequencies at that site. The method uses a Bayesian approach to infer the parameters of the model and employs sparse shrinkage to account for the fact that only a small number of motifs affect the mutation rate.

Context-Dependent Base Alteration Model

For the sake of simplicity, we hereby refer to a base alteration as a mutation. This is often convenient because a base alteration may be captured in NGS experiments as an observed mutation (e.g., adenine to inosine alterations are observed as adenine to guanine mutations following sequencing [Levanon et al. 2004]). Let the full sequence (e.g., the genome) we are interested in be denoted as $G = (g_1, \ldots, g_n)$, where $g_i \in \{A, C, G, T$ and $|G| = n$. Note that upon analyzing RNA genomes T is replaced by U, however, for convenience, we keep the notation of T through the text. We denote the k-long context of a focal position, i, as the $\lfloor \frac{k}{2} \rfloor$ nucleotides flanking the position, that is, the sequence $g_{i-\lfloor \frac{k}{2} \rfloor}, \ldots, g_i, \ldots, g_{i+\lfloor \frac{k}{2} \rfloor}$, where k is assumed to be an odd number. For example, for $k = 3$ and $G = AGGAT$ there are three distinct contexts, AGG, GGA, and GAT (fig. 1).

We can further decompose sequence contexts into motifs (which will be the features in our feature selection algorithm). In the example above, a possible motif would be AGX (where X is any of the four nucleotides). For $k = 1$, there are four motifs (the four possible nucleotides), for $k = 3$ there are 3×4 motifs consisting of one nucleotide, $\binom{3}{2} 4^2$ motifs consisting of two nucleotides, and 4^3 motif consisting of three

nucleotides. In general, for k-long contexts, there are $m = \sum_{i=1}^{k} \binom{k}{i} 4^i$ motifs (fig. 1 and table 1). Next, we consider a mutation operating on a context, $\mu = (a \rightarrow b)$, to be an ordered pair, where $a, b \in \{A, C, G, T\}$, and $a \neq b$. For simplicity, we denote this as a context + mutation (fig. 1).

Now, let $\boldsymbol{X}^{(k)}_{c \times m}$ be a matrix that indicates, for each context + mutation, the motifs it contains such that:

$$x_{ij} = \begin{cases} 1 & \text{the } i\text{th context contains the } j\text{th motif} \\ 0 & \text{otherwise} \end{cases},$$

where c is the number of all unique contexts + mutations in genome G· c is at most 3×4^k (because not all contexts may be present in a given genome), and m is defined above as the number of possible motifs (table 1).

We assume sequencing of a population of homologous genomes and the availability of a known reference genome (fig. 2). These assumptions allow us to uniquely define which mutation occurred at what sequence context. We can define the vector $\vec{y}$ to be the empirical count of contexts + mutations for each type of mutation. We define $\vec{l}$ to be a vector of the total observed number of occurrences of each context in the sequencing data (so that $\vec{l}_i$ is the sequencing coverage of the context). In other words, we pool all the mutations that have the exact same context across different positions in the genome so that in figure 2 we would count three mutations for the red context. Note that when using counts we implicitly assume lack of genetic drift which may have increased the copy number of an allele in the population. Alternatively, the input data may be the number of polymorphic loci with a specific context, that is, a position is then counted at most once. Accordingly, in figure 2 we would count two polymorphisms for the red context. The former representation fits our viral data set; however, the latter enables flexibility for other data sets as well, supporting the method's general applicability. Table 1 illustrates an example showing X, $\vec{y}$, and $\vec{l}$ for a context of size $k = 3$ for simplicity, using the pooled mutation approach.

Model

We introduce CIPI—context-based inference of point-mutation influence. We will use a logistic regression model with the latent variables $\alpha \in R$, $\sigma \in R^+$, $\vec{\beta} \in R^m$, and $\vec{\gamma} \in \{0, 1^m$ (where m is the number of all possible motifs), which relates motifs in a context to the probability of observing a mutation in the context, defined as $\vec{p^{\mu}} \in R^c$, where μ was referred earlier as a mutation type. We assume $p_i^{\mu} = \text{logit}^{-1}(\sum_{j=1}^{m} \beta_j \gamma_j x_{ij} + \alpha)$. For convenience, from now on we will denote p_i^{μ} simply as p_i given that the inference is performed per mutation type. The introduction of the $\vec{\gamma}$ indicator vector allows us to better penalize models with a large number of influencing motifs, that is, it will allow us to use shrinkage in the model to avoid overfitting. We further introduce κ, a factor that can control for shrinkage that is elaborated on below. Thus, our logistic regression model is defined by $\vec{\gamma}$, a vector of indicator variables indicating whether the ith variable contributes significantly to the model or not, $\vec{\beta}$, the vector of regression coefficients that increase or decrease the probability of a mutation in a context, α, the baseline mutation rate in the absence of any context, and κ, a shrinkage factor. We use the properties of the $\text{logit}^{-1}(x) = \frac{\exp(x)}{\exp(x)+1}$ which maps $\text{logit} : R \rightarrow (0, 1)$ which guarantees that p_i is always a valid probability.

FIG. 2. Population mutations as a factor of the genomic contexts in the ancestral genome. We start with an ancestral genome (also referred to here as a reference genome). In this original genome there are several sequence contexts, three of which are illustrated here in colored boxes; the red context, the blue context, and the yellow context. A context might be present in the genome once (e.g., blue and yellow contexts), might have multiple appearances (the red context), or not to be present at all. After t generations the population is no longer homogenous, and mutations might arise in different contexts (colored x marks). Elevated mutation rate or genetic drift may lead to more mutations at a specific locus (see text for more details).

Inference

We will use Bayesian variable selection to estimate $\vec{\gamma}$. Our approach is similar to classical methods used for QTL mapping. These methods aim to identify correlations between a set of genetic markers (e.g., single nucleotide polymorphisms) and a continuous phenotype (Yi 2004). There is usually a very large number of single nucleotide polymorphisms (which take the role of features) in a sample and a much smaller number of samples. Moreover, often the different features are correlated, mainly due to linkage in a genetic cross or linkage disequilibrium in outbred populations. In our case, the motifs are the features and the number of possible motifs is often much larger than the number of mutations observed in a given genome, especially when considering microbial genomes that tend to be relatively small. Furthermore, there is also a strong correlation between the different motifs (features), as these may be nested within each other or may be mutually exclusive. We address these problems by assigning a prior distribution to each latent variable. The model is too complex to allow analytical solutions, but we can infer the posterior distribution using a Markov chain Monte Carlo (MCMC) algorithm.

Posterior Probability Calculations

The posterior probability of the four latent variables can be written as:

Table 1. An Example of Sequencing Data and Associated Breakdown into Context, Motifs, and Counts.

	Context and mutation	AXX	XXT	...	TXT	$\vec{y}$ (counts of all mutations with a given context)	$\vec{l}$ (number of observations at the given context)
		← m →					
c	$A(C\rightarrow T)A$	1	0	...	0	110	20000
	$T(C\rightarrow T)T$	0	1	...	1	84	202
	...						

NOTE.—The main variables, X, are shaded in gray, representing the indicators for presence/absence of motifs within a context. In the first row of this example, we observe C→T mutations within the context ACA in 110 of the individuals in the population sequenced, out of a total of 20,000 sequences (reads) covering all ACA contexts.

$$\mathrm{p}(\vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa | X, \vec{y}, \vec{l}) = \frac{\mathrm{p}(X, \vec{y}, \vec{l} | \vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa) \times \mathrm{p}(\vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa)}{\sum_{\gamma', \beta', \alpha', \sigma', \kappa'} \mathrm{p}(X, \vec{y}, \vec{l} | \vec{\gamma}', \vec{\beta}', \alpha', \sigma', \kappa') \times \mathrm{p}(\vec{\gamma}', \vec{\beta}', \alpha', \sigma', \kappa')} \propto$$

$$p(X, \vec{y}, \vec{l} | \vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa) \times \mathrm{p}(\vec{\gamma}, \vec{\beta}, \kappa)$$

$$= p(X, \vec{y}, \vec{l} | \vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa) p(\vec{\beta} | \vec{\gamma}, \sigma) \mathrm{p}(\vec{\gamma} | \kappa) \times p(\kappa) p(\alpha) p(\sigma).$$

The expressions in the numerator can easily be computed when values for the latent variables are given: $p(X, \vec{y}, \vec{l} | \vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa) = \prod_{i=1}^{c} \binom{l_i}{y_i} p_i^{y_i} (1-p_i)^{l_i - y_i}$ where p_i is defined above, c is defined above as the number of context + mutations, and we assume that $y_i \sim$ Binomial (l_i, p_i). Notice that we assume here that the number of mutations observed for each context is small so that different mutations in the same context are approximately independent of each other. On the other hand, calculating the sum in the denominator is intractable for all possible combinations of $\vec{\gamma}, \vec{\beta}, \alpha, \sigma, \kappa$.

Prior Probability Specification

Prior for κ

κ is the model shrinkage factor controlling the model's sparsity. Small κ values will result in no signal at all (shrinkage is too high); however, larger κ values will increase the model complexity. We define a uniform prior on κ, such that $\kappa \sim U(10^{-200}, 10^{-2})$. The boundaries were set in order to allow high shrinkage.

Prior for $\vec{\gamma}$

We define a simple Bernoulli prior probability distribution for each element of $\vec{\gamma}$ so that $\Pr(\vec{\gamma} | \kappa) = \kappa^q \times (1-\kappa)^{m-q}$, where q is the number of "1" entries in $\vec{\gamma}$. Taking $\kappa = 0.5$ defines a prior that gives equal weight to any $\vec{\gamma}$. In practical estimation, the collinearity between the predictors can lead to instability. The collinearity for this problem has been described previously as a "dilution" effect (George 2010). For example, if $XXCGX$ is a feature that increases the mutation rate for C⟶T mutations, naïve estimation of effects by counting might also find that $AXCGX$ significantly increases the mutation rates, as the $AXCGX$ motif contains the $XXCGX$ motif. To address this problem, we add a penalty to the prior for $\vec{\gamma}$. We use a method based on determinantal point process (DPP) that has been shown to be effective in other Bayesian variable selection problems (Ročková and George 2014; Kojima and Komaki 2014, 2016). According to the DPP method, the prior is weighted by powers of the determinant of the correlation matrix, $\Pr(\vec{\gamma} | \kappa) \propto \left| X_\gamma^T X_\gamma \right|^w \kappa^q \times (1-\kappa)^{m-q}$, where X_γ is a $c \times q$ matrix including only the columns $\{i | \vec{\gamma}_i = 1\}$ from the original matrix X and $w \in R^+$ is a weight factor. The weighting provides a computationally tractable approach for mitigating the effect of the dependencies between the features. If all features are completely independent $\left| X_\gamma^T X_\gamma \right| = 1$, whereas in the case of full dependency (collinearity) between at least one pair of vectors the matrix will be singular and $\left| X_\gamma^T X_\gamma \right| = 0$.

Prior for β

We define the prior probability for the regression coefficients, $\vec{\beta} | \vec{\gamma}, \sigma$, as:

$$p(\vec{\beta} | \vec{\gamma}) = \prod_{j=1}^{m} p(\beta_j | \gamma_j = 0)^{I(\gamma_j = 0)} p(\beta_j | \gamma_j = 1)^{I(\gamma_j = 1)},$$

Where $I(\cdot)$ is the indicator function, $p(\beta_j | \gamma_j = 0) = N(0, \sigma^2)$, and $p(\beta_j | \gamma_j = 1) = N(0, C^2 \sigma^2)$ for some variance σ^2 and some constant C. It is possible to either infer σ^2 from the data or to define it as a constant; here we chose to infer it from the data. Notably when $\gamma_j = 0$, the regression coefficient is undefined and β_j is unidentifiable.

Prior for α

α represents the mean rate of base modification, which we refer to here as the base mutation rate. We assume α is normally distributed $p(\alpha) = N(r_{\text{mean}}, r_{\text{SD}}^2)$ where $r_i = \frac{y_i}{l_i}$ for $i \in \{1 \ldots c\}$, and use the empirical mean and

variance inferred from the data, thus $r_{\text{mean}} = \frac{\Sigma_i r_i}{m}$, $r_{SD} = \sqrt{\frac{\Sigma_i (r_{\text{mean}} - r_i)^2}{m-1}}$.

Prior for σ^2

σ^2 can be set as a hyperparameter or can be inferred from the data with a variety of possible priors. A simple prior for σ^2 is the uniform prior $\sigma^2 \sim U(\text{low}, \text{high})$, where "low" and "high" are set arbitrarily.

MCMC Implementation

To traverse the posterior distribution, we chose to use MCMC using a Metropolis–Hastings method (Hastings 1970).

Our MCMC algorithm works as follows:

(1) We start from an arbitrary point $(\beta_0, \gamma_0, \alpha_0, \sigma_0, \kappa_0)$.

(2) We then define a transition kernel, that is, the set of probabilities for proposing a new parameter value given the previous value. Let the parameters in the ith step be $\beta_i, \gamma_i, \alpha_i, \sigma_i, \kappa_i$. A new set of parameters $\beta_i^*, \alpha_i^*, \sigma_i^*$ are sampled based on Gaussian distributions for each variable: $\beta_i^* \sim N(\beta_i, I_m)$, $\alpha_i^* \sim N(\alpha_i, 1)$, $\sigma_i^* \sim N(\sigma_i, 1)$. For γ_i^* the update kernel assumes $|\gamma_i - \gamma_i^*| = d \sim Pois(\lambda)$ for some λ and randomly choose d entries to flip. For κ_i^* we sample based on a Gaussian distribution; $\kappa_i^* \sim N(\kappa_i, \min(10^{-1}, \kappa_i \cdot 10^{2\cdot\lambda}))$, where λ is the number of accepted steps in the previous 100 iterations, doubled since there are two possible directions (i.e.,increasing or decreasing κ). As we aim for lower kappa values,we set a lower bound on the distribution's standard deviation. The use of $\kappa_i \cdot 10^{2\cdot\lambda}$ will result in bigger steps upon extensive acceptance, and smaller steps towards convergence.

Then we accept $\beta_i^*, \gamma_i^*, \alpha_i^*, \sigma_i^*, \kappa_i^*$ with probability of $\min\left(1, \frac{Pr(X, \vec{y}, \vec{l} | \beta_i^*, \gamma_i^*, \alpha_i^*, \sigma_i^*, \kappa_i^*) \pi(\beta_i^*, \gamma_i^*, \alpha_i^*, \sigma_i^*, \kappa_i^*)}{Pr(X, \vec{y}, \vec{l} | \beta_i, \gamma_i, \alpha_i, \sigma_i, \kappa_i) \pi(\beta_i, \gamma_i, \alpha_i, \sigma_i, \kappa_i)}\right)$ where $\pi(\cdot)$ is the prior.

(3) After the chain has completed a predetermined number of iterations (burn-in), we use the ergodic averages of each parameter to approximate the posteriors.

Results

Simulated Data Sets

In order to verify the performance of our method, we simulated population NGS data. Our aim was to mimic evolving populations of viruses where rare mutations are often observed widely across the genome. Accordingly, our simulations mimicked population sequencing of oral poliovirus (Stern et al. 2017) with a genome length of ~7,500 bases at a sequencing depth of ~100,000 reads per locus. Using an adaptation of the Moran model (Materials and Methods), we introduced an increased or reduced mutation frequency based on $k = 5$ contexts. We assume a given ancestral sequence and simulate each position independently, in line with our short-term evolutionary experiments, in which we do not expect more than one mutation per genome over the course of the experiment. Fourteen generations of mutations only (no selection) were simulated, with a population size of $N = 10^5$ and mutation rate of $u = 10^{-5}$. In each of 500 simulated data sets, we introduced a different number (between 0 and 3) of motifs influencing the mutation rate of a specific mutation type. We then used our inference framework to infer which were the influential motifs in each of the simulations. We defined a threshold, t, such that if the posterior probability that motif i has an effect on the mutation rate is larger than t, $p(\gamma_i = 1 | X, \vec{y}, \vec{l}) > t$, we predict that motif to affect the mutation rate. The posterior probability of $\gamma_i = 1$ is estimated as the average occupancy time of the Markov chain in the state $\gamma_i = 1$.

Figure 3 summarizes the results of the simulations for given different thresholds of t. We will consider a motif as a true positive if it was originally simulated and determined as having an effect on mutation rate for a given threshold t, and consider a motif as a false positive if it was not originally simulated but detected as having such effect. A motif that was not simulated and was not determined as effecting mutation rates will be considered as a true negative. All in all, the accuracy rate was very high across all thresholds tested, demonstrating the power of the approach to accurately detect the influence of a motif on the rate of mutation. High accuracy is contributed due to remarkably high true negative rate (specificity) as expected from imbalanced data predictions. This means that given a set of all possible motifs for a chosen k, the number of influencing motifs which satisfy $p(\gamma_i = 1 | X, \vec{y}, \vec{l}) > t$ for a large enough t will be miniscule compared with the noninfluencing ones. For example, a model that predicts $\gamma_i = 0$ for all i will naturally result in high true negative rate (most motifs are eventually labeled as 0) and consequently high accuracy. Thus, we choose to use precision and recall as our evaluation metrics. We obtained the precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$), also known as sensitivity and used the F1 score, which is the harmonic mean of precision and recall ($2 \cdot \frac{p \cdot r}{p+r}$), to evaluate the trade-off between precision and recall.

In order to approximate the false discovery rate (FDR), we measured the Bayesian FDR defined as the expected probability of labeling a motif as nonsignificant given that the motif is considered significant by a threshold t. More formally:

$$\text{Bayesian FDR} = E\left[P\left(\gamma_i = 0 | X, \vec{y}, \vec{l}, \bar{\gamma}_i \geq t\right)\right] \text{ for all } i$$

Where $\bar{\gamma}_i$ is the ergodic average calculated to approximate the posterior as defined in the MCMC implementation.

Given a threshold of 0.8, our false positive rate is <3%, suggesting that the method has high specificity, that is, we correctly reject noninfluential motifs (fig. 3). The low false positive rate promotes high precision; however, not all motifs are successfully identified, leading to lower recall. When analyzing the few false positives of the method, we noticed that the vast majority of motifs incorrectly inferred were motifs embedded in the true simulated motif (e.g., AXXXX when the true motif was AAXXX). This suggests that the DPP method

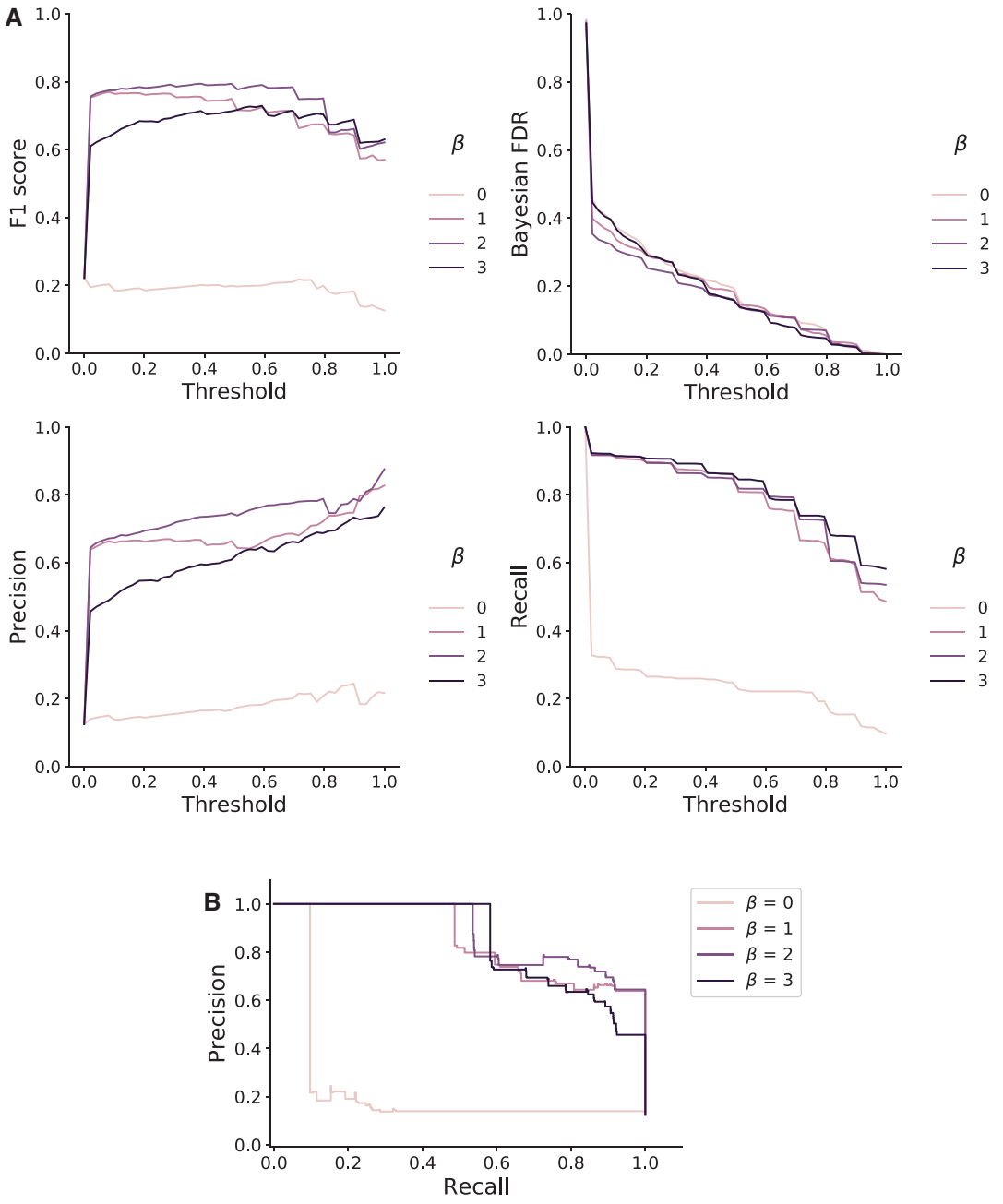

FIG. 3. Accuracy analysis of CIPI on simulated data, as a factor of β values. (*A*) From left to right: F1 score, Bayesian FDR, precision, and recall. For $\beta = 0$, the F1 score is substantially low as we impose zero contribution to all motifs. Increasing beta highly improves our prediction, converging almost to the same scores. (*B*) Precision–recall curve.

used to remove correlated motifs still has its limitations, and the incorrectly inferred motifs are still valid (see supplementary fig. S1, Supplementary Material online).

Introducing Noise into the Analysis

In the simulations above, all sites were simulated as neutral, and the observed frequencies were assumed to be the true frequencies. However, in real biological data, these two assumptions will likely be problematic. For one, there is no set of genomic sites known to be completely neutral. Typically, synonymous sites are assumed to be neutral, yet a subset of these sites may be under selection, particularly in viruses (Chamary et al. 2006). Moreover, the observed mutation frequencies will be affected by sampling and thus will likely deviate from true mutation frequencies. Thus, in order to test how our method fares with noisier data, we tested the inference of CIPI on "noisy" data, defined here as such when (i) a proportion of sites deviate from neutrality, and (ii) sampling is applied as described above. The accuracy of inference was examined as a function of β, the proportion of nonneutral sites, and sequencing noise. Setting β to zero yields no context effect, as $\sum_{j=1}^{m} \beta_j \gamma_j X_{Qj} + \alpha = \alpha$, thus we expect the F1 score to asymptotically approach zero. We show that for ranging βs the effects of noise are negligible.

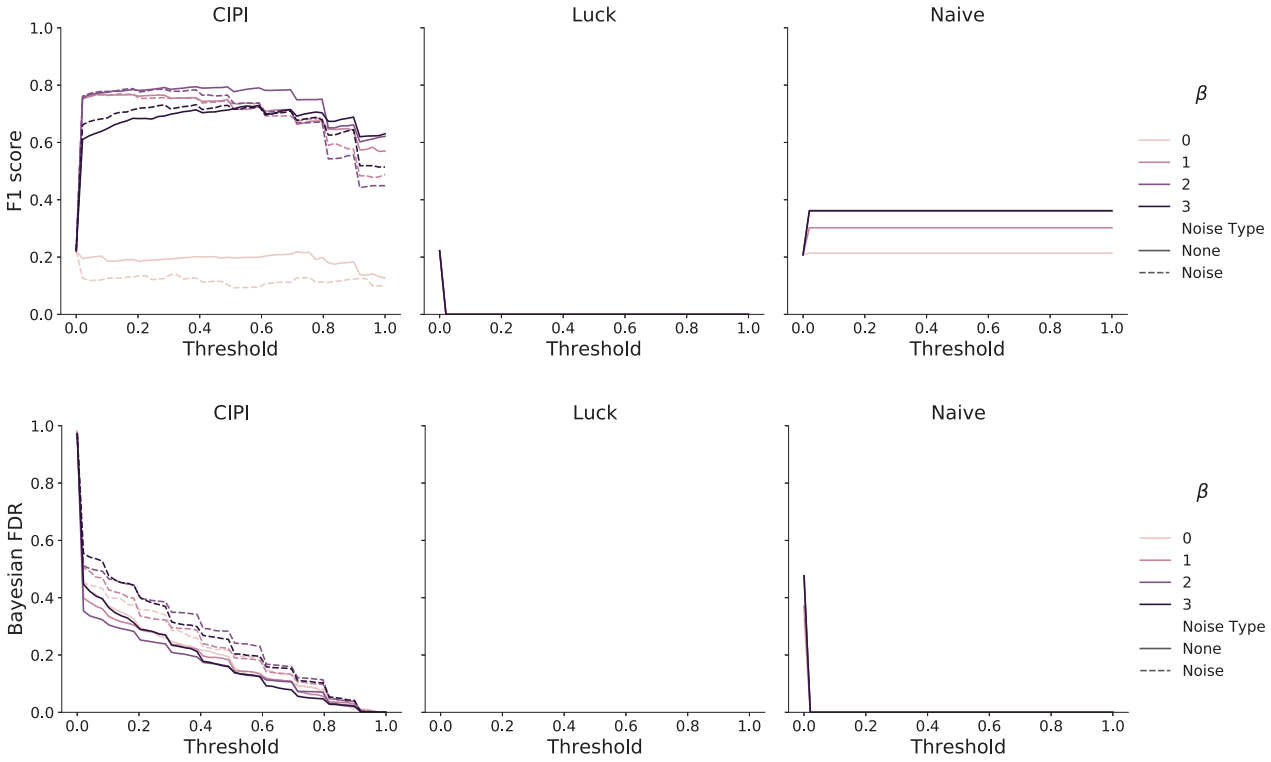

FIG. 4. Performance comparison for analysis of noisy and non-noisy simulated data sets. F1 scores (upper panel) and Bayesian FDR (lower panel) are shown for nonnoisy (solid lines) and noisy data (dashed lines) for our CIPI model (left panels). Mid and right panels, respectively, demonstrate reduced performance of a random model (Luck) and a naive model (main text).

We were interested in comparing our method to alternative methods. However, many methods are based on assumptions that are violated in our case, mostly the lack of a phylogenetic tree for NGS data (Siepel and Haussler 2003) or the fact we are dealing with short genomes where not all contexts exist (Aggarwala and Voight 2016). We thus implemented a naïve method for inferring sequence context, where the motifs that have the highest mutation rate are picked (Materials and Methods). Notably variations of this naïve method are widely employed in the literature today (Schneider and Stephens 1990; Sandelin et al. 2004; Dey et al. 2018). We also compared our results with a model based on "random" inference, which assigns $\gamma_i = 0$ for all i. Our results show dramatically superior inference to both the naïve and random approaches (fig. 4). (For precision and recall as a factor of threshold and performance under extreme selection see supplementary figs. S2 and S3, Supplementary Material online.)

Oral Polio Vaccine Virus Experimental Evolution Data Set

To further examine the biological applicability of the method, we applied it to oral poliovirus 2 (OPV2) sequencing data (Stern et al. 2017). The experiment was designed to record the mutation frequencies of OPV as it was serially passaged in tissue culture. Notably, the very high sequencing depth of this experiment, spanning $10^5 - 10^6$ reads per position, combined with very high mutation rates of the virus (spanning $10^{-4} - 10^{-5}$ mutations per base per replication cycle), made these data perfect for our model. In order to rule out the effects of selection that can easily mimic the effects of increased or decreased mutation rate, our analysis focused only on synonymous mutations that are mostly (although not always) neutral. We further focused only on transition mutations, as transversions are less frequent and hence inferred with less reliability. We analyzed the last and seventh passage, which corresponded to 14 viral replication cycles.

Table 2 presents the resulting motifs detected in passage 7, at a threshold of $t = 0.8$ for the probability that the motif affects mutation rates. Intriguingly, many of the motifs detected are compatible with editing by the enzymes ADAR1 or ADAR2 (Eggington et al. 2011). Both enzymes edit adenosine to inosine, which is detected as a A→G mutation, and prefer A or U (T) upstream to the edited A. As polioviruses copy both the positive and negative RNA strand syntheses in the cell (Schulte et al. 2015), a T→C mutation on the reverse-complement negative strand will be read as an A→G on the positive strand (which is the reference genome against which all reads are mapped). Accordingly, table 2 shows enrichment for A→G and T→C mutations as compared with the composition of the OPV2 genome ($P < 0.001$, Fisher exact test).

APOBEC Signatures Discovered in HIV-1 Data Set

To demonstrate the applicability of the method on different types of NGS data, we obtained sequencing data of HIV-1 knowing to contain strong APOBEC3G signatures (Pollpeter et al. 2018). APOBEC3G (A3G) is an antiviral host factor with

Table 2. Motifs Affecting Mutation Rates Detected in Empirical Data of Passage 7 of OPV2.

Mutation Type	Motif	Reverse Complement Motif	Mean Gamma	Mean Mutation Rate	Context Mean Mutation Rate	Increase/ Decrease
A→G	XXAGA	TCTXX	1	0.00041	0.00079	Increase
A→G	**XTACX**	**XGTAX**	**1**	**0.00041**	**0.00113**	**Increase**
A→G	XXAAG	CTTXX	1	0.00041	0.00098	Increase
A→G	CXAGX	XCTXG	1	0.00041	0.00019	Decrease
A→G	**XAAGX**	**XXTTT**	**1**	**0.00041**	**0.00113**	**Increase**
A→G	XXAGG	CCTXX	1	0.00041	0.00021	Decrease
C→T	GCCXX	XXGGC	1	0.00072	0.00027	Decrease
C→T	CXCCX	XGGXG	1	0.00072	0.00025	Decrease
G→A	XXGCX	XGCXX	1	0.00015	0.00085	Increase
G→A	XCGXX	XXCGX	1	0.00015	0.00085	Increase
T→C	**TXTAX**	**XTAXA**	**1**	**0.00048**	**0.0012**	**Increase**
T→C	AXTXG	CXAXT	1	0.00048	0.0015	Increase
T→C	**XXTAA**	**TTAXX**	**1**	**0.00048**	**0.0008**	**Increase**
T→C	XCTXG	CXAGX	0.95	0.00048	0.00016	Decrease
T→C	**XGTTX**	**XAACX**	**0.99**	**0.00048**	**0.00021**	**Decrease**
T→C	XTTXA	TXAAX	0.83	0.00048	0.0003	Decrease

NOTE.—Motifs exceeding a threshold (mean gamma) of 0.8 are presented. Motifs in the context of ADAR are shown in bold.

Table 3. Motifs with Positive Effect on Transition Mutation Rates Detected in HIV-1 Empirical Data.

Mutation Type	Motif	Mean Gamma	Mean Mutation Rate	Context Mean Mutation Rate	Increase/Decrease
G→A	AGGXX	1	0.08	0.22	Increase
G→A	TTGXX	0.93	0.08	0.1	Increase
G→A	XTGXT	1	0.08	0.14	Increase
G→A	**CXGGX**	**1**	**0.08**	**0.43**	**Increase**
G→A	**TXGGX**	**1**	**0.08**	**0.22**	**Increase**
G→A	**XXGGG**	**1**	**0.08**	**0.59**	**Increase**
G→A	**XXGAG**	**1**	**0.08**	**0.37**	**Increase**

NOTE.—Motifs exceeding a threshold (mean gamma) of 0.8 are presented. All presented motifs are of mutation type G→A. The transitions C→T, T→C, and A→G did not contain motifs which were significant for positive effects. The motifs in A3G context are shown in bold.

cytidine deaminase activity, which in our terms is a C →U (T) mutation. A3G was shown to deaminate the nascent minus DNA strand of HIV-1, leading to an observed G→A hypermutation on the +RNA genomic strand (Harris et al. 2003). The DNA editing function of A3G is known to be context dependent, where a focal G is observed to be followed by a "G" or an "A" (G**G** or G**A**). We set out to analyze data from an experiment that measured A3G activity directly on nascent DNA strands (Pollpeter et al. 2018), and ran our inference with $k = 5$. Out of seven significant motifs, four were related to the A3G context with $\beta \sim 3$, suggesting that we are also capable of recognizing known enzymatic signatures on empirical data with strong selection effects. The results are presented in table 3.

Running Times and Resources

We report running times of our method, obtained for a 2.5 GHz Intel Core i7 processor with 16 Gb of RAM. The analysis ran with a constant rate of 400 iterations per second, and running 10^6 steps in one chain required ~50 min for a ~7,000-base-long genome and utilized ~300 MB of available RAM.

Discussion

We developed here a novel approach for the detection and evaluation of sequence context on mutation rates. Our prime motivation was to develop a method that is able to analyze high-resolution deep sequencing data sets. One of the main challenges in these data sets is the high dimensionality of the data when accounting for sequence context. Thus, one of the main novelties of the new approach is the use of Bayesian shrinkage to take into account the fact that the number of sequence motifs that affect mutation rates is likely much smaller than the number of possible motifs. We conclude that the method provides highly accurate results on simulated data. In particular, we precisely identify the motifs which do not influence mutation rate. Our remarkably low false positive rate promotes high confidence in inferring influencing motifs; however, we also fail to detect some true motifs, suggesting that our method is conservative. We believe that this trend is due to a combination of the combinatorial complexity of the inference and the Bayesian shrinkage leading to a sparse $\vec{\gamma}$ vector. Too strong shrinkage might cause the dilution of an existing effect, as setting the initial $\vec{\gamma}$ vector to zero will show no effect at all.

Our analysis of empirical poliovirus data revealed the potential effect of ADAR, a known protein known to edit virus genomes. Remarkably, we were able to detect this effect despite very low mutation frequencies of ADAR-associated motifs (table 2). Further experimental work is required to validate this finding. We also demonstrated that our method fares well with other types of empirical

data where the effects of selection are much stronger. We successfully captured motifs which are associated with APOBEC3G editing in HIV-1 data, and thus we conclude that the method can be generally applicable to virus NGS data.

Having said that, we would like to recognize some limitations that could restrict the reliability of inferences based on our method. The assumption of neutrality is essential to the analysis as point mutations which strongly deviate from neutrality might promote the appearance of a context effect. This might be especially true for empirical data. The more data that are available (e.g., the more synonymous sites analyzed), the more the effect of a handful of nonneutral sites will be diminished.

In the analysis described here, we assumed the absence of genetic drift by pooling counts of all mutations in a given context. Our experiments were performed using very large population sizes; however, the copy number of the mutations was very low, and thus genetic drift is most likely very prominent, mostly in the earlier passages of the experiment. To this end we focused on the last passage where we noted fewer fluctuations of mutation frequencies, associated with drift (Zinger et al. 2019). Although random genetic drift is not directly modeled, our simulations do incorporate genetic drift. Reassuringly, we did not observe false positives due to drift, most likely because drift will not affect one specific context consistently. We note that our model does allow analysis of data where one counts polymorphisms rather than mutations; for our setup where high mutation rates of viruses lead to a polymorphism at most sites of the genome, this is irrelevant.

To summarize, we have developed a robust method that has the potential and the strength to identify influential sequence contexts, and this may shed light upon the underlying mechanisms of both polymerases and other enzymes that render genetic modifications. The method is flexible, compatible with a wide variety of applications and data sets and should be fairly easily executed for the analysis of NGS data. Although the method was designed with our recent experiments of virus populations in mind, it is also generic enough for other types of NGS data.

The python code for the method is publicly available via GitHub under https://github.com/SternLabTAU/CIPI. The data used in this study are available through Zenodo under the https://zenodo.org/record/3408598 with a direct link through the GitHub repository.

Materials and Methods

Moran Model for Simulating Context Dependence

To simulate viral sequences where the rate of mutation depends on the context, we use an adaptation of the four-allelic Moran model previously described for simulating the evolution of cancer cells (Zhu et al. 2011). This model is a continuous time birth–death model. Usefully, this model takes into account large population sizes and allows for different mutation rates, both of which are relevant in our case. We simulate each position independently, while taking into account the context of the position as described below. Each allele is one of the four nucleotides.

Let N be the population size. For a given position i, at time t we define a vector $V^t \in N^4$ $s.t$ $\Sigma_j V_j^t = N$. Thus, each entry V_j^t is the number of genomes with the jth allele in the ith position. We initialize

$$V_j^0 = \begin{cases} N & \text{if the original nucleotide at position } i \text{ is } j \\ 0 & \text{else} \end{cases}$$

meaning that we start the simulations from a homogenous population of genomes defined by the original sequence. When an individual of type j dies it is replaced by randomly selecting a parent from the N options and copying it. Notably, the model we used (Zhu et al. 2011) also defines, for each phenotype j, a different fitness value $(1 + s_j)$ and different mutation rates $\mu_{jj'}$ for each $j \neq j'$ that changes the allele of the individual. In the first set of simulations reported in the Results section, we assume neutrality for all mutations, and hence set $s_j = 0$ for every j. In order to model the influence of context, we assume that there are different $\mu_{jj'}$ for different sequence contexts. For each position at a given context Q, $\mu_{jj'}^Q = \text{logit}^{-1}(\sum_{j=1}^m \beta_j \gamma_j X_{Qj} + \alpha)$, similar to the way context was assumed to affect mutation rates in our model above. The evolutionary process defined above is a Markov chain and we assume that births/deaths occur at a constant rate in time such that time can be measured continuously. At time t the state of the chain is $V^t = (V_1^t, V_2^t, V_3^t, V_4^t)$. For a four-allelic model, we have 12 possible events that can change the state of the chain, $V^t \rightarrow V^{t'}$, by decreasing the value of V_j^t and increasing the value of $V_{j'}^t$ for some values of j and j', $j \neq j'$.

The rate of such a transition is $r(V^t, V^{t\prime}) = \frac{V_j^t \times V_{j'}^t (1+s_{j'})}{\sum_i V_i^t (1+s_i)} + V_j^t \cdot \mu_{jj'}^Q$.

Let $R = \begin{pmatrix} r_1 \\ \dots \\ r_{12} \end{pmatrix}$ be the rates of all 12 possible transitions that can occur at any given point in time. Furthermore, define a matrix $S_{4\times 12}$ where entry i, j of the matrix equals

$$S_{i,j} = \begin{cases} 1 & \text{transition } j \text{ increments nucleotide } i \text{ by } 1 \\ -1 & \text{transition } j \text{ decreases nucleotide } i \text{ by } 1 \\ 0 & \text{else} \end{cases}.$$

Then, the change in allele frequency through time is given by $\frac{dV}{dt} = S \times R$. Using standard stochastic simulation algorithms (Gillespie 1977), we then simulate this process to generate samples of genetic data. As the mutations are all assumed to be segregating at low frequencies, which is true for the experimental viral sequence evolution data that we hope to emulate, we will approximate the sequence evolution as the union of multiple independently evolving sequence sites.

Simulating Noisy Frequencies and the Effects of Purifying Selection

In order to simulate mutation frequencies that underwent sampling, we began with the set of simulated frequencies described above which we assumed to be our true frequencies and assumed $N = 1{,}000{,}000$ genomes. We then applied binomial sampling on the frequencies and sampled 100,000 genomes for sequencing. We further tested the method when a proportion of sites deviate from neutrality. We simulated 500 data sets with a varying number of loci under selection, where selection coefficients were sampled based on a distribution of fitness effects from an empirical data set of RNA viruses (Sanjuan et al. 2004).

Comparison with a Naïve Approach

We wished to compare our method with other approaches. However, all existing methods were inappropriate for our data: We cannot use any phylogenetic methods as the short read NGS data do not allow reconstructing a phylogenetic tree. Also, we cannot use the regression analysis described in Aggarwala and Voight (2016), which originally was used for analyzing the human genome, as the transition into polymorphisms in a small genome violates several of the model's assumptions. Thus, we set out to implement a naïve solution for inferring the motifs in which the mutation rate is elevated. Given the frequencies of all mutations in a genome and a defined mutation type, we considered the top 1% contexts with the highest average mutation rates. Formally, for a context C that appears n times in a genome we define $f_\mu^C(i)$ as the frequency of the mutation $\mu = (a \to b)$ at position i in the context C. Then, we average all frequencies of all positions that share the same context.

The average mutation rate in context C will be:

$$\bar{f}_\mu^C = \frac{1}{n}\sum_{i=1}^{n} f_\mu^C(i).$$

For the top 1% most frequent contexts, we obtain all the contained motifs and label them as significant.

Empirical Data analysis—Poliovirus Data Set

We analyzed the oral poliovirus 2 (OPV2) sequencing data reported by (Stern et al. 2017), using mutation frequencies as reported therein.

Empirical Data analysis—HIV Data Set

We analyzed HIV-1 data reported by Pollpeter et al. (2018). Raw fastq files were downloaded and mapped using an in-house computational pipeline to the HIV-1 reference genome pNL4-3 (accession number AF324493.2) at positions 1–180, as per the original paper. The coverage obtained matched the paper's description.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Karin Mittelman and Shiran Abadi for their constructive suggestions during the writing of this paper. This study was supported in part by a fellowship to G.L. and D.M. from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, by funding from the Israeli Science Foundation (1333/16) to A.S., by funding from Raymond and Beverly Sackler Fund to A.S. and R.N., and by funding by the Koret-UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics to A.S. and R.N.

References

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

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet*. 48(4):349–355.

Aikens RC, Johnson KE, Voight BF. 2019. Signals of variation in human mutation rate at multiple levels of sequence context. *Mol Biol Evol*. 36(5):955–965.

Berikov VB, Rogozin IB. 1999. Regression trees for analysis of mutational spectra in nucleotide sequences. *Bioinformatics* 15(7):553–562.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*. 7(2):98–108.

Cooper DN, Krawczak M. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet*. 83(2):181–188.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* 274(5673):775–780.

Cuevas JM, Sanjuán R, Garijo R, López-Aldeguer J, Geller R. 2015. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol*. 13(9):e1002251.

Dey KK, Xie D, Stephens M. 2018. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinformatics* 19(1):473.

Dunson DB, Tindall KR. 2000. Bayesian analysis of mutational spectra. *Genetics* 156(3):1411–1418.

Eggington JM, Greene T, Bass BL. 2011. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun*. 2:319.

Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. 2006. DNA motifs associated with aberrant CpG island methylation. *Genomics* 87(5):572–579.

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol*. 33(1):268–280.

George EI. 2010. IMS collections borrowing strength: theory powering applications – a festschrift for dilution priors: compensating for model space redundancy. *Inst Math Stat*. 6:158–165.

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 81(25):2340–2361.

Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *Elife* 6.

Harris RS, Sheehy AM, Craig HM, Malim MH, Neuberger MS. 2003. DNA deamination: not just a trigger for antibody diversification but also a mechanism for defense against retroviruses. *Nat Immunol*. 4(7):641–643.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol Biol Evol*. 24(10):2196–2202.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. 12(11):756–766.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*. 101(39):13994–14001.

Kojima M, Komaki F. 2016. Determinantal point process priors for Bayesian variable selection in linear regression. *Stat Sin*. 26:97–117.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet*. 63(2):474–488.

Lehmann KA, Bass BL. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry* 39(42):12875–12884.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol*. 22(8):1001–1005.

Lunter G, Hein J. 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 20(Suppl 1):i216–i223.

Narayan P, Ludwiczak RL, Goodwin EC, Rottman FM. 1994. Context effects on N6-adenosine methylation sites in prolactin mRNA. *Nucleic Acids Res*. 22(3):419–426.

Pollpeter D, Parsons M, Sobala AE, Coxhead S, Lang RD, Bruns AM, Papaioannou S, McDonnell JM, Apolonia L, Chowdhury JA, et al. 2018. Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G. *Nat Microbiol*. 3(2):220–233.

Razin A, Riggs AD. 1980. DNA methylation and gene function. *Science* 210(4470):604–610.

Ročková V, George EI. 2014. Determinantal priors for variable selection. 47th Scientific Meeting of Italian Statistical Society.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol*. 26(7):1663–1676.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*. 107(10):4629–4634.

Rogozin IB, Malyarchuk BA, Pavlov YI, Milanesil L. 2005. From context-dependence of mutations to molecular mechanisms of mutagenesis. *Pac Symp Biocomput*. 2005:409–20.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 32(90001):D91–D94.

Sanjuan R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A*. 101(22):8396–8401.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. 18(20):6097–6100.

Schulte MB, Draghi JA, Plotkin JB, Andino R. 2015. Experimentally guided models reveal replication principles that shape the mutation distribution of RNA viruses. *Elife* 4:1–18.

Siepel A, Haussler D. 2003. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*. 21(3):468–488.

Simmonds P, Xia W, Baillie J, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* 14(1):610.

Stern A, Yeh MT, Zinger T, Smith M, Wright C, Ling G, Nielsen R, Macadam A, Andino R, Yoneyama T, et al. 2017. The evolutionary pathway to virulence of an RNA virus. *Cell* 169(1):35–46.e19.

Sung W, Ackerman MS, Gout J-F, Miller SF, Williams E, Foster PL, Lynch M. 2015. Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. *Mol Biol Evol*. 32(7):1672–1683.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. 2013. ADAR regulates RNA editing, transcript stability, and gene expression. *Cell Rep*. 5(3):849–860.

Wulff BE, Sakurai M, Nishikura K. 2011. Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat Rev Genet*. 12(2):81–85.

Yi N. 2004. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* 167(2):967–975.

Zhang W, Bouffard GG, Wallace SS, Bond JP, Program N. 2007. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. *J Mol Evol*. 65(3):207–214.

Zhu T, Hu Y, Ma ZM, Zhang DX, Li T, Yang Z. 2011. Efficient simulation under a population genetics model of carcinogenesis. *Bioinformatics* 27(6):837–843.

Zhu Y, Neeman T, Yap VB, Huttley GA. 2017. Statistical methods for identifying sequence motifs affecting point mutations. *Genetics* 205(2):843–856.

Zinger T, Gelbart M, Miller D, Pennings PS, Stern A. 2019. Inferring population genetics parameters of evolving viruses using time-series data. *Virus Evol*. 2019;5(1):vez011.