JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 2, 2013

Pp. 113–123
DOI: 10.1089/cmb.2012.0232

Probabilistic Inference of Viral Quasispecies Subject to Recombination

ARMIN TÖPFER,[1,2,*] OSVALDO ZAGORDI,[3,*] SANDHYA PRABHAKARAN,[4] VOLKER ROTH,[4] ERAN HALPERIN,[5,6] and NIKO BEERENWINKEL[1,2]

ABSTRACT

RNA viruses exist in their hosts as populations of different but related strains. The virus population, often called quasispecies, is shaped by a combination of genetic change and natural selection. Genetic change is due to both point mutations and recombination events. We present a jumping hidden Markov model that describes the generation of viral quasispecies and a method to infer its parameters from next-generation sequencing data. The model introduces position-specific probability tables over the sequence alphabet to explain the diversity that can be found in the population at each site. Recombination events are indicated by a change of state, allowing a single observed read to originate from multiple sequences. We present a specific implementation of the expectation maximization (EM) algorithm to find maximum *a posteriori* estimates of the model parameters and a method to estimate the distribution of viral strains in the quasispecies. The model is validated on simulated data, showing the advantage of explicitly taking the recombination process into account, and applied to reads obtained from a clinical HIV sample.

Key words: evolution, HMM, statistical models, viruses.

1. INTRODUCTION

Next-generation sequencing (NGS) technologies have transformed experiments previously considered too labor-intensive into routine tasks (Metzker, 2010). One application of NGS is the sequencing of genetically heterogeneous populations to quantify their genetic diversity. The genetic diversity is of primary interest, for example, in infection by RNA viruses such as HIV and HCV. In these systems, the combination of a high mutation rate of the pathogen with recombination between pathogens gives rise to a population of different but related individuals, referred to as a viral quasispecies. Recombination can occur,

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.
[2]SIB Swiss Institute of Bioinformatics, Basel, Switzerland.
[3]Institute of Medical Virology, University of Zurich, Zurich, Switzerland.
[4]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland.
[5]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel.
[6]International Computer Science Institute, Berkeley, California.
*These authors contributed equally to this work.

as a subsequent event, when different viral particles infect a single cell. We denote the different viral strains in this population as haplotypes. Studying the features of the viral quasispecies can shed light on the mechanisms of pathogen evolution in the host and it is of direct clinical relevance. For example, the diversity of the quasispecies has been shown to affect virulence (Vignuzzi et al., 2006), immune escape (Nowak et al., 1991), and drug resistance (Johnson et al., 2008).

The quasispecies equation is a mathematical model for RNA virus populations evolving according to a mutation-selection process (Eigen, 1971). The dynamics of the model are described by a mutation term accounting for transformation of one viral haplotype (or strain) into another at the time of replication, and a selection term that accounts for varying replication rates of different strains. The mutation process is generally considered as the result of point mutations only, although recombination is known to be frequent in many clinically relevant viruses, including HIV and HCV. For example, the recombination rate of HIV is estimated to be about tenfold higher than its point mutation rate. Therefore, the quasispecies model has been extended to account for both mutation and recombination (Boerlijst et al., 1996). At equilibrium, the model predicts the viral population to be dominated by one or a few haplotypes, which are surrounded by a cloud of constantly generated, low-frequency mutants.

Recent NGS technologies allow for observing viral quasispecies at an unprecedented level of detail by producing millions of DNA reads in a single experiment. However, this high yield comes at a cost. Reads are usually short, up to 700 bp with the latest technology and much shorter than the smallest viral genomes, and they are error prone due to sample preparation and sequencing errors (Gilles et al., 2011). As a result, since the data obtained are incomplete and noisy, a meaningful characterization of viral populations by means of NGS requires careful analysis of the sequencing data (Beerenwinkel and Zagordi, 2011).

In this article, we aim at inferring viral quasispecies based on NGS data by explicitly modeling mutation and recombination. We use a hidden Markov model (HMM) to generate viral populations, i.e., haplotype distributions, and their probing by means of NGS. In our model, the haplotypes are originating from a small number of generating sequences via recombination, described as switch of state in the HMM that selects from which sequence the haplotype derives, and mutation, described by position-specific probability tables for the generating sequences. The sequencing reads are obtained from the haplotypes subject to observation error.

HMMs allowing for a switch between generating sequences, termed jumping HMMs, have been applied, for example, to sequence alignment of protein domains (Spang et al., 2002) and to detecting inter-host HIV-circulating recombinant forms (Schultz et al, 2006). A related model has been used in human genetics to infer the haplotypes of diploid genomes from genotype data (Kimmel and Shamir, 2005; Scheet and Stephens, 2006). The model presented here differs from previous approaches in several ways, including an unknown and possibly large number of haplotypes, erroneous sequence read data, and high mutation and recombination rates. In particular, our goal is to reliably identify the haplotypes shaping intra-host quasispecies, including variants of low frequencies. Since sequencing errors will confound the true variation present in the sample, methods for error correction have been proposed, including clustering of reads or flowgrams followed by removal of any remaining within-cluster variation (Eriksson et al., 2008; Quince et al., 2011; Zagordi et al., 2010a, 2010b).

In the present article, we present a novel generative probabilistic model for making inference of viral quasispecies, i.e., for estimating the intra-patient viral haplotype distribution. Specifically, we assume that the true genetic diversity is generated by a few sequences, called generators, through mutation and recombination, and that the observed diversity results from additional sequencing errors. We present the model for local haplotype inference, meaning that we aim at inferring the population structure in a genomic region of a size that can be covered by individual reads, but extending this model to global haplotype inference, i.e., to longer genomic regions, is straightforward. Local inference will generally be more reliable and sufficient for many applications. For example, the HIV protease gene, an important target of antiretroviral therapy, is 297 bp long, and it is now standard to obtain reads of 400 bp and longer with the Roche/454 GS Junior sequencer, a common pyrosequencing platform for clinical diagnostics. Local haplotype reconstructions can also be used as a starting point for global reconstruction (Astrovskaya, 2011; Eriksson et al., 2008; Prabhakaran et al., 2010; Prosperi et al., 2011; Zagordi et al., 2011). We show that our model is able to estimate the distribution of viral haplotypes with high reliability by applying it to simulated data, where we have access to the ground truth, and we present an application to reads obtained from an HIV-infected patient.

2. METHODS

2.1. Hidden Markov model

During infection of a host cell, a viral strain can change either by point mutation, when a single base is copied with error, or by recombination, when a cell is infected by more than one viral particle, and viruses in subsequent generations produce a sequence that is a mosaic of those of the progenitors. The model we present here does not aim at representing these evolutionary processes mechanistically. Rather, it is a descriptive probabilistic model, in which the quasispecies is generated by switching among K different generating sequences, each of length L. We denote by π_k the probability to begin with generator k at the first sequence position. The generators are defined as sequence profiles (μ_{jkv}), indicating the probability over the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, -\}$ of base $v \in \mathcal{A}$ at position j of the k-th generating sequence.

The set of sequences generates viral haplotypes $H \in \mathcal{A}^L$ by mutation, modeled by the probability tables (μ_{jkv}), and recombination, denoted by transition matrices ρ_j. The transition probability ρ_{jkl} describes the recombination event in which the generating sequence k switches to l between positions $j-1$ and j.

Let Z_j be the hidden random variable with state space $[K] = \{1, \ldots, K\}$, indicating the parental sequence generating H_j, the haplotype character at position j. Each observed read R with bases R_j is obtained from a haplotype subject to noise (sequencing errors), assumed to occur independently among sites at rates ε_j. The probability of an observed read R is defined hierarchically as

$$\Pr(Z_1 = k) = \pi_k \tag{1a}$$

$$\Pr(Z_j = l \mid Z_{j-1} = k) = \rho_{jkl} \tag{1b}$$

$$\Pr(H_j = v \mid Z_j = k) = \mu_{jkv} \tag{1c}$$

$$\Pr(R_j = b \mid H_j = v) = \begin{cases} \varepsilon_j & \text{if } b \neq v \\ 1 - (n-1)\varepsilon_j & \text{otherwise} \end{cases} \tag{1d}$$

where $n = |\mathcal{A}|$ is the size of the alphabet.

The full model consists, for each observation $i = 1, \ldots, N$, of the hidden random variables Z_j^i, indicating generator sequences, and H_j^i, the haplotypes of the quasispecies, and the observed reads R_j^i, for all sequence positions $j = 1, \ldots, L$ (Fig. 1). The model parameters are summarized as $\theta = (\pi, \rho, \mu, \varepsilon)$.

For parameter estimation, we first describe the maximum likelihood (ML) approach and develop an Expectation Maximization (EM) algorithm for ML estimation (MLE). Then, we define prior parameter distributions that enforce sparse maximum *a posteriori* (MAP) solutions and present a modified EM algorithm for MAP estimation of the parameters. This regularized model is used subsequently for the rest of the article.

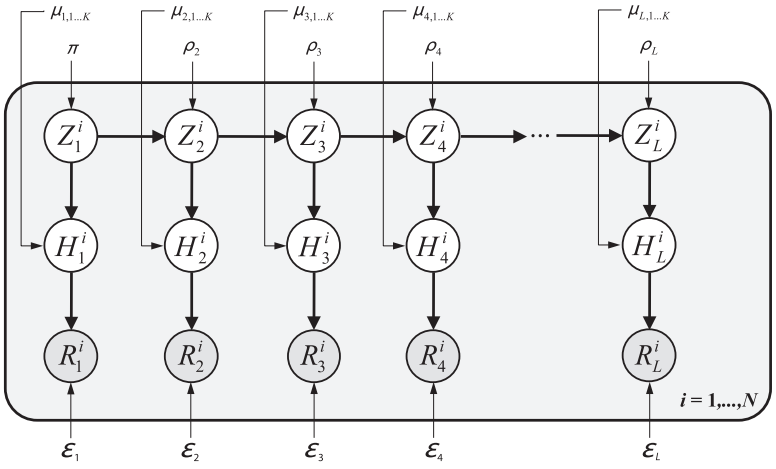

FIG. 1. Graphical representation of the model. Only one observation i is depicted; for the full model, the graph is replicated for $i = 1, \ldots, N$.

2.2. *Maximum likelihood estimation*

The likelihood $\Pr(R \mid \theta)$ of the model defined in Eqs. 1a–d factorizes into the product over independent reads, and for each read, it can be computed efficiently using the Markov property,

$$\begin{aligned}\Pr(R \mid \theta) &= \prod_i \sum_{Z^i, H^i} \Pr(Z^i, H^i, R^i) \\ &= \prod_i \sum_{Z^i, H^i} \prod_j \Pr(R^i_j \mid H^i_j)\Pr(H^i_j \mid Z^i_j)\Pr(Z^i_j \mid Z^i_{j-1}),\end{aligned}$$

where $\Pr(Z^i_1 \mid Z^i_0) = \Pr(Z^i_1)$. Using the distributive law, each sum in this expression can be factored along the Markov chain, which gives rise to the forward algorithm (Rabiner, 1989). In this manner, the likelihood can be computed in $O(NLK^2)$ time.

The EM algorithm (Dempster, 1977) is an iterative procedure to find local maxima of the likelihood as a function of θ by maximizing the auxiliary Baum's function $Q(\theta, \theta')$, defined as the expected hidden log-likelihood of the data with respect to the posterior distribution of (Z, H) given θ',

$$Q(\theta, \theta') = \mathrm{E}_{Z, H \mid \theta'}[\log \Pr(R, Z, H \mid \theta)].$$

Here, θ' is the previous estimate of the parameters $(\pi, \rho, \mu, \varepsilon)$. Baum's function bounds the log-likelihood from below, and repeated iterations of maximizing Q with respect to θ (M-step) alternated with estimation of the posterior $\Pr(Z, H \mid R, \theta)$ (E-step) are guaranteed to find a local maximum of the likelihood function.

For the E-step, we compute

$$\begin{aligned}Q(\theta, \theta') = &\sum_k N^{\mathrm{jump}}_1(k) \log \pi_k + \sum_{j=2}^{L} \sum_{k,l} N^{\mathrm{jump}}_j(k, l) \log \rho_{jkl} + \sum_{j,k,v} N^{\mathrm{hap}}_j(k, v) \log \mu_{jkv} \\ &+ \sum_{j,v} N^{\mathrm{read}}_j(v, v) \log (1 - (n-1)\varepsilon_j) + \sum_{j, v \neq b} N^{\mathrm{read}}_j(v, b) \log \varepsilon_j,\end{aligned}$$

where $N^{\mathrm{jump}}_1(k)$ is the expected number of times a Markov chain starts in state k at position 1, $N^{\mathrm{jump}}_j(k, l)$ is the expected number of times that a Markov chain switches from state k to state l right before position j, $N^{\mathrm{hap}}_j(k, v)$ is the expected number of times the Markov chain is in state k and emits haplotype character v at position j, and $N^{\mathrm{read}}_j(v, b)$ is the expected number of times the Markov chain emits character v at position j and character b is observed in the reads. These expected counts are estimated for all reads by computing posterior probabilities of the hidden variables H and Z given the data and the current estimate of θ, using the forward and backward algorithm (Rabiner, 1989).

In the M-step, the parameters are updated by maximizing $Q(\theta, \theta')$ with respect to θ. This is achieved by setting

$$\pi_k = \frac{N^{\mathrm{jump}}_1(k)}{\sum_{k'} N^{\mathrm{jump}}_1(k')}, \qquad \rho_{jkl} = \frac{N^{\mathrm{jump}}_j(k, l)}{\sum_{l'} N^{\mathrm{jump}}_j(k, l')},$$

$$\mu_{jkv} = \frac{N^{\mathrm{hap}}_j(k, v)}{\sum_{v'} N^{\mathrm{hap}}_j(k, v')}, \qquad \varepsilon_j = \frac{\sum_{v \neq b} N^{\mathrm{read}}_j(v, b)}{N(n-1)}.$$

2.3. *Maximum* a posteriori *estimation*

The HMM defined by Eqs. 1a–d is non-identifiable (Ito et al., 1992). In this case, multiple solutions of θ share the same posterior. Hence, MLEs are not uniquely defined and the EM algorithm suffers from poor convergence. To address this limitation, we define a prior distribution for the model parameters and estimate them by maximizing the posterior probability

$$\Pr(\theta \mid R) \propto \Pr(R \mid \theta)\Pr(\theta).$$

We assume independent priors, $\Pr(\theta) = \Pr(\pi)\Pr(\rho)\Pr(\mu)\Pr(\varepsilon)$, and $\Pr(\pi)$, and $\Pr(\varepsilon)$ to be flat. For the recombination probabilities ρ and the nucleotide probability tables μ, we define independent and identical Dirichlet distributions for all sequence positions j and all generators k,

$$\rho_{jk} \sim \mathrm{Dir}(\alpha_0, \ldots, \alpha_0),$$
$$\mu_{jk} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_1).$$

The hyperparameter α_0 controls the sparsity of recombination events. As α_0 approaches zero, the transition matrix ρ_j, at the MAP estimate, approaches the $K \times K$ identity matrix I_K, and recombination becomes more unlikely. Similarly, α_1 controls the variability of haplotype character emissions. For small values of α_1, mutations become increasingly unlikely. The Dirichlet priors can enforce sparse MAP solutions and for a high degree of regularization (small α_0 and α_1), the model becomes identifiable.

The regularized model is not only computationally more convenient, but sparse recombination is also a biologically plausible assumption. Indeed, despite high recombination rates, real RNA virus populations always display genomic regions that are conserved or nearly conserved, and that define the virus. In these regions, the different generating sequences cannot be distinguished because there is no or little diversity. Therefore, recombination among different sequences can only be observed in regions with higher diversity, which are a small fraction of genomic sites, and thus is expected to be a rare event.

For solving the MAP estimation problem, we use the Variational Bayes approach suggested and elaborated in Beal (2003) and Johnson (2007). With the Dirichlet priors defined above, it can be solved by a modification of the EM algorithm introduced in the previous section. Specifically, only the M-step needs to be modified to update ρ and μ as follows:

$$\rho_{jkl} \propto \frac{f(N_j^{\mathrm{jump}}(k, l) + \alpha_0)}{f(\sum_{l'} N_j^{\mathrm{jump}}(k, l') + K\alpha_0)},$$
$$\mu_{jkv} \propto \frac{f(N_j^{\mathrm{hap}}(k, v) + \alpha_1)}{f(\sum_{v'} N_j^{\mathrm{hap}}(k, v') + n\alpha_1)},$$

where the scaling function $f(x) = e^{\psi(x)}$ is defined in terms of the digamma function ψ, the derivative of the log gamma function, and the constants of proportionality are given by the constraints $\sum_l \rho_{jkl} = 1$ and $\sum_v \mu_{jkv} = 1$, respectively.

2.4. Model selection

For fixed sequence alphabet $\mathcal{A}$ and length L, the dimension of the model is determined by the number K of generator sequences. For model selection, i.e., for choosing the optimal K, we consider the Bayesian information criterion defined as

$$\log \Pr(R \mid \hat{\theta}_{\mathrm{MAP}}) - \frac{\nu \log N}{2},$$

where ν is the dimension, or number of free parameters, of the model (Schwarz, 1978). However, the size of the model makes it infeasible to compute its dimension directly by standard methods. Instead, we resort to a heuristic inspired by the results in (Yamazaki, 2005) and approximate ν by the number of nonzero parameters in the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$. In our empirical tests, this heuristic worked very well for values of K up to at least 5. For practical purposes, this appears sufficient, as in the quasispecies model the number of dominant haplotypes is assumed to be small and hence can be generated by a small number K of low-entropy probability tables over $\mathcal{A}$. For model selection, the smallest K is chosen within one standard error of the K with the maximum BIC (Tibshirani et al., 2001).

As an alternative model selection strategy, the goodness of fit of the model may be assessed in a cross-validation setting, but this approach is computationally much more expensive and would drastically slow down the effective runtime.

2.5. Prediction

The main object of interest that we derive from the model is the haplotype distribution $\Pr(H)$, i.e., the structure of the viral quasispecies. For given model parameters θ, estimated from read data R, we can compute

the probability of each haplotype efficiently using the forward algorithm, and the distribution $\Pr(H)$ might be estimated by computing the probability of each haplotype. However, since there are 4^L possible haplotypes, enumeration is infeasible already for moderate sequence lengths L. Instead, we estimate $\Pr(H)$ by sampling from the model using Eqs. 1a–c. In the applications below, we sampled 10,000 haplotypes at the MAP estimate of θ. This procedure is efficient, because the entropy of almost all model parameter probability tables μ_{jk} and ρ_{jk} is close to zero, and hence the probability mass of $\Pr(H)$ will be centered on a few haplotypes.

Although not employed in the present article, other quantities of interest can be predicted from the model. For example, read error correction can be done by replacing each read R^i with the haplotype it most probably originated from, i.e., with $\operatorname{argmax}_h \Pr(H^i = h \mid R^i = r)$. If the allele frequency spectrum is sought after, for example, because the effect of specific single-nucleotide variants (SNVs) is known, then the posterior probability of each SNV given all observed reads can be computed as $\Pr(H_j = v \mid R)$.

2.6. Implementation

For the EM algorithm, we iterated the E-step and the M-step until convergence, which was detected by a relative change of the log-likelihood smaller than 10^{-8}. Since the EM algorithm is only guaranteed to find a local maximum, we performed 50 random restarts and chose the solution with the largest likelihood.

The initial parameter values for μ_{jk} are constant, $\mu_{jk} = (1/n, \ldots, 1/n)$, and for π and ρ_{jk}, they are drawn at random from the distributions $\pi \sim \mathrm{Dir}(2, \ldots, 2)$ and $\rho_{jk} \sim \mathrm{Dir}(\tau_0, \ldots, \tau_0)$, where τ_0 controls the sparsity. If necessary, the values of ρ_{jk} are reordered such that the k-th entry always has the highest value of this vector, which is achieved by switching two entries. The resulting initial transition matrix encodes rare transitions across the generators. Empirical analysis indicated good performance of MAP estimation with prior and initialization Dirichlet hyperparameters set to $\alpha_0 = \alpha_1 = \tau_0 = 0.01$.

The forward and backward calculations of the E-step will underflow for long reads; therefore, these probabilities are rescaled at each step rather than computed on a logarithmic scale, which would take much more time. Reads are hashed at the beginning in order to identify identical ones and to avoid unnecessary computations. Thus, the effective runtime is $O(N_u L K^2)$, where N_u is the number of unique reads. The E-step can be independently computed for each read. In addition, all random restarts of the EM can be computed separately.

We have implemented MAP estimation, model selection, and prediction of the haplotype distribution in a Java program called QuasiRecomb. It runs on any operating system supporting Java version 1.7 (Linux, OSX, Solaris, Windows). The software is open source and licensed under the GNU General Public License. It is available online at `www.cbg.ethz.ch/software/quasirecomb/`

3. RESULTS

3.1. Simulation study

We assessed the performance of our model on seven different datasets, corresponding to different distributions of haplotypes of 300 bp length. Datasets 1, 2, 3, 5, and 6 have one recombination breakpoint, dataset 4 has two recombination breakpoints, and dataset C serves as a negative control without any recombinants. Generator sequences differed by between 6 and 10% of nucleotides. The haplotype distribution of each dataset is reported in Figure 2.

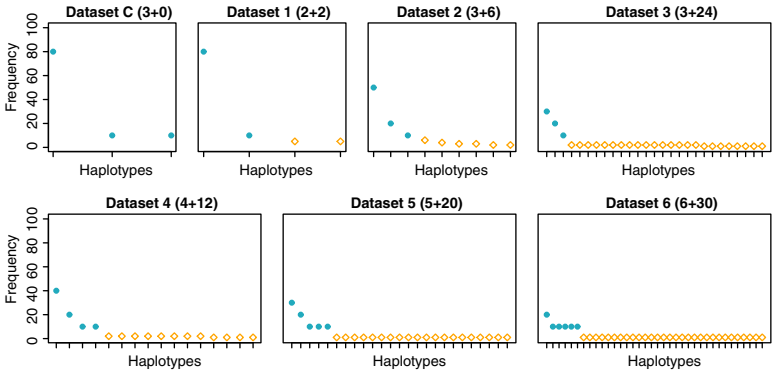

FIG. 2. Frequencies in percent for the haplotypes of each dataset. The symbols for the original haplotypes and recombinants are filled dots and open squares, respectively. The numbers in parentheses on top of each plot report the number of original haplotypes plus the number of recombinants in the respective dataset.

For each distribution of haplotypes, we sampled 50 datasets of 2,000 reads each with point mutations at an error rate of 0.03% per base and evaluated the BIC score for model selection. The error rate reflects the amount of substitution errors that can be expected in a typical NGS experiment using 454/Roche after filtering low-quality reads and removing frameshift-causing indels in the alignment step. Figure 3 reports the BIC scores for the seven datasets. In all cases, the correct number of generators is chosen, applying our model selection. Except for the last dataset, the BIC score is maximum at the correct number of generators.

In order to study the impact of the sample size, we sampled instances of the first dataset of different sizes and repeated the model selection procedure. We analyzed datasets of 300, 400, and 500 reads. The results are reported in Figure 4. For the first two samples, the BIC score erroneously selects $K = 1$, whereas for 500 and more reads, the procedure correctly selects $K = 2$ generating sequences, indicating that sufficient coverage is an important prerequisite.

Furthermore, the sensitivity has been tested using the first dataset, with fixed haplotype distances. We sampled 1,000 error-free reads from the original haplotypes at frequencies 80% and 20%, and replaced arbitrarily two reads of the sample with the two recombinants from the first dataset. QuasiRecomb was able to reconstruct these recombinants at a frequency of 0.1%. Since the scenario of a technical error-free sample is unrealistic, we again sampled as before, but with an error-rate of 0.03% per base, and again we are able to reconstruct the low-frequency recombinants. Even in the case of a tenfold higher error-rate of 0.3% per base, QuasiRecomb successfully identifies and reconstructs the recombinants. This can be explained by the fact that an accumulation of about 10 technical errors on a length of 300 bp is very unlikely. One might also investigate the sensitivity w.r.t. the distances among haplotypes and fixed frequencies.

For parameter estimation, we sampled an additional set of 2,000 reads from the haplotype distribution and ran the EM algorithm with the value of K inferred before. Then, we inspected the MAP estimates of the parameters μ, π, and ρ. Whenever the correct K had been chosen, the entropy of the estimates of the tables μ_{jk} is very close to zero. Regarding the recombination parameters, except for the negative control dataset, there is always a position j after the last variable site and before the recombination hotspot such that $\rho_{jkl} \neq 0$ for two different generators k and l.

In the negative control dataset, where no recombinants are present, the three generating sequence profiles corresponded exactly to the original haplotypes, i.e., the entropy of all μ_{jk} tables was very close to zero, and no recombination was detected, i.e., $\rho_j = I_K$ for all j. In this case, π represents the frequency of the original haplotypes and its estimate was very close to the original distribution. The remaining discrepancy can be explained by the sampling variance of the reads alone.

We assessed the accuracy of the inferred quasispecies by comparing it to the original set of sequences using the proportion close measure, φ_q, defined as the fraction of reconstructed haplotypes that match an original one with at most q mismatches (Eriksson, 2008). Figure 5 reports the proportion close, as a function of the number of allowed mismatches, for the models learned from datasets 1 to 4. For comparison, we additionally learned a model in which recombination is not possible, i.e., where $\rho_j = I_K$ for all j.

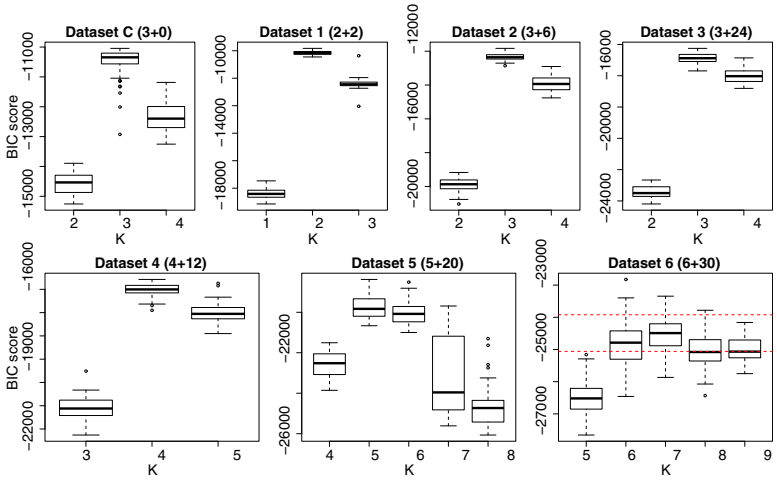

FIG. 3. BIC score for the seven simulated datasets. The model correctly chooses K for up to five original haplotypes. The boxplots summarize results of 50 independent datasets. The numbers in parentheses report the number of original generators plus the number of recombinants. The maximum BIC is attained at the true number K of generators for all datasets except #6, where the maximum BIC is at $K = 7$, but our model selection selects the true $K = 6$ (dashed lines report upper and lower boundaries of the one standard error of $K = 7$).

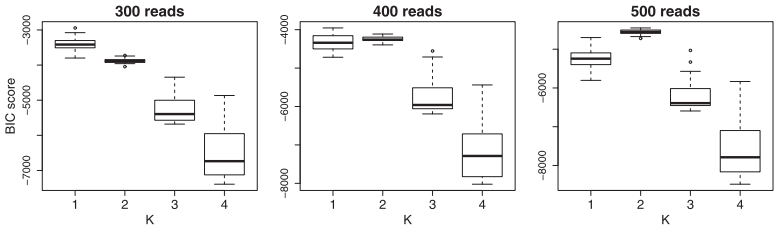

FIG. 4. BIC score for simulated dataset 1 at different sample sizes between 300 and 500 reads. The model selection correctly selects $K = 2$ already with 500 reads. The boxplots summarize results on 50 independent datasets.

Allowing recombination, model selection chose the correct number of generators for all four datasets. Without recombination, the correct number of generators (and of haplotypes) is $4 = 2 + 2$, $9 = 3 + 6$, $27 = 3 + 24$, and $16 = 4 + 12$, respectively (Fig. 3). Identifying nine generators or more is hard, and running the EM algorithm becomes inefficient, because the runtime grows quadratically in K. For dataset 1, without recombination, BIC still selects $K = 2$, because the penalization for a larger K is too high. Visual inspection of the MAP estimate parameters shows that the first generator is completely concentrated on one profile, and the second generator explains the other three haplotypes by flat generator distributions μ_{j2}. Setting $K = 4$, all generators concentrate on the original haplotypes, but as expected, the runtime is higher and many more EM restarts are needed to find this MAP estimate, because the likelihood surface is very flat.

The advantage of modeling recombinants is evident as the fraction of the population reconstructed is always higher than in the recombinant-free case. For all four datasets, the proportion close is at least 99% for $q \geq 1$ if recombination is accounted for, whereas the recombination-free model fails to reconstruct the quasispecies structure in these cases (Fig. 5). This is a consequence of the poor performance of the model selection in this case, namely $K = 2$, 4, 5, and 5 for datasets 1, 2, 3, and 4, respectively.

3.2. *Real HIV dataset*

Using QuasiRecomb, we analyzed a set of experimental NGS reads obtained by sequencing a clinical sample from an HIV-infected patient in the context of a study of viral tropism (Archer et al., 2010) (Sequence Read Archive run SRR069887). We selected 1,517 reads overlapping a 179 bp long region of the *env* gene (positions 6321–6499 in the HXB2 reference strain). We ran the EM algorithm on 50 datasets, generated by bootstrapping 1,517 reads each, and selected the model with $K = 2$ generators (Fig. 6, left). The estimated quasispecies is dominated by a single haplotype with an estimated frequency of 31%. This master sequence is surrounded by a swarm of mutants, 12 of which have relative frequencies over 1% and many others have lower frequencies (Fig. 6, right). The sequence similarities of the estimated haplotypes with relative frequencies greater than 1% are between 93 and 99%. Each generator has four positions with a positive recombination probability.

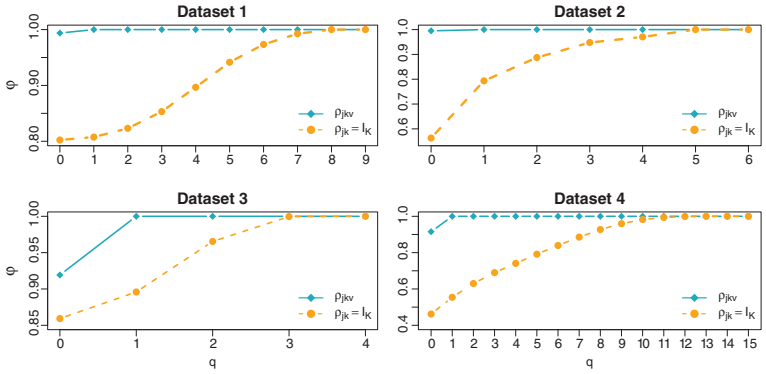

FIG. 5. Proportion close, φ_q, as a function of q. The fraction of the population reconstructed with at most q mismatches is higher than 99% already for $q \geq 1$ if one allows recombination, but only for $q = 3$ to 15 if one does not allow recombination.

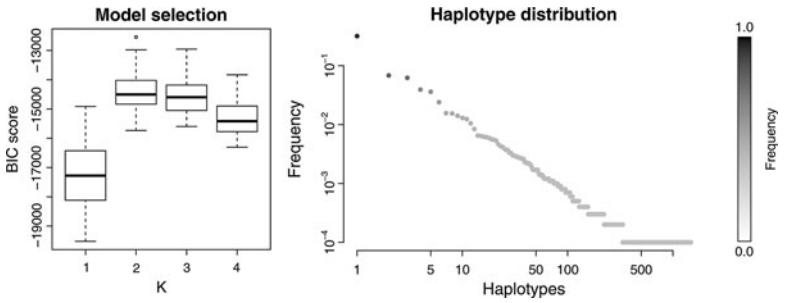

FIG. 6. BIC scores for a clinical sample based on 50 bootstrap samples (left) and the haplotype distribution of the inferred quasispecies on a log–log scale for $K = 2$ (right).

In order to appreciate the compactness of the model inferred with the jumping HMM, we compared its solutions with those of another tool to reconstruct haplotypes, implemented in the software ShoRAH (Zagordi, 2011). This method, which does not take recombination into account, identified 15 haplotypes in the same dataset, which can be further reduced to 10 if one excludes those with frequencies lower than 1% and those which harbor a frameshift due to a deletion.

4. DISCUSSION

We have presented a probabilistic model based on an HMM that infers the distribution of haplotypes in a viral quasispecies from NGS data. The model describes these different viral strains present in the population as originating from different generating sequences by means of two processes: point mutation and recombination. Point mutation is captured by the fact that the sequences are modeled as probability tables over the sequence alphabet. Recombination is modeled via a change of the sequence from which the haplotype is drawn, as indicated by a change of state, or a jump, in the hidden Markov chain. Due to the possibility of switching between sequences, the number of tables necessary to describe the population structure remains small, while offering an excellent fit to the data. This results in a more compact and structured description of the viral population.

We have introduced regularization to achieve sparse MAP estimates accounting for the fact that mutation and recombination are rare events. Using the EM algorithm, MAP estimates can be computed efficiently. Our results on simulated data demonstrate the usefulness of enforcing this sparsity when inferring recombinant haplotypes from read data.

There are several ways to extend the methodology presented here. Estimating the haplotype distribution is currently done by sampling from the model. Another approach would be to compute the top suboptimal haplotypes from the recombinant sequence generators. In previous work on the analysis of NGS data to estimate genetic diversity, model selection has been approached in a non-parametric way by using the Dirichlet process mixture (Prabhakaran et al., 2010; Zagordi, 2010a). Extension of the HMM in this direction has been proposed and might be explored in this context as well (Beal et al., 2002).

We have presented our results in a local reconstruction setting, but our implementation QuasiRecomb is already adapted to accept global read alignments in BAM format. In this scenario, the population structure inferred locally is extended to genomic regions that are longer than the typical read length. This is achieved by allowing for longer generating sequences, along with two additional silent states, to describe the unobserved regions before and after each read in the same fashion as the pair-HMM can be used for semi-global sequence alignment.

The quasispecies inference approach we propose here is designed for sequencing technologies with long reads. In general, we expect the accuracy of our method to decrease for shorter reads. We note, however, that the read length of most sequencing technologies is constantly improving and that some NGS platforms can produce reads over 1,000 bp. With such long reads, the probability to observe recombinations on a single read will be higher, and the necessity to keep the number of generators small will be even more compelling.

ACKNOWLEDGMENT

Eran Halperin is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. He was supported by the Israeli Science Foundation (grant 04514831).

DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Archer, J., Rambaut, A., Taillon, B.E., et al. 2010. The evolutionary analysis of emerging low frequency hiv-1 cxcr4 using variants through time—an ultra-deep approach. PLoS Comput. Biol. 6, e1001022.

Astrovskaya, I., Tork, B., Mangul, S., et al. 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12.

Beal, M., Ghahramani, Z., Rasmussen, C. 2002. The infinite hidden Markov model. Advances in neural information 14, 577–584.

Beal, M.J. 2003. Variational algorithms for approximate bayesian inference [Tech. rep.,] University College London.

Beerenwinkel, N., and Zagordi, O. 2011. Ultra-deep sequencing for the analysis of viral populations. Current Opinion in Virology 1, 413–418.

Boerlijst, M., Bonhoeffer, S., and Nowak, M. 1996. Viral quasi-species and recombination. Proceedings: Biological Sciences 263, 1577–1584.

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussions). J. R. Statist. Soc. B. 39, 1–38.

Eigen, M. 1971. Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften. Available online at www.springerlink.com/index/q47866457218x543.pdf

Eriksson, N., Pachter, L., Mitsuya, Y., et al. 2008. Viral population estimation using pyrosequencing. PLoS Computational Biology 4, e1000074.

Gilles, A., Meglécz, E., Pech, N., et al., 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12, 245.

Ito, H., Amari, S.I., and Kobayashi, K. 1992 Identifiability of hidden Markov information sources and their minimum degrees of freedom. IEEE Transactions on Information Theory 38, 324—333.

Johnson, J.A., Li, J.F., Wei, X., et al. 2008. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. Plos Med. e158.

Johnson, M. 2007. Why doesn't EM find good HMM POS-taggers? EMNLP-CoNLL, 296–305. www.aclweb.org/anthology/D/D07/D07-1031

Kimmel, G., and Shamir, R. 2005 GERBIL: Genotype resolution and block identification using likelihood. Proc. Natl. Acad. Sci. U. S. A. 102, 158–62.

Metzker, M.L. 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. 11, 31–46.

Nowak, M.A., Anderson, R.M., McLean, A.R., et al. 1991. Antigenic diversity thresholds and the development of AIDS. Science 254, 963–9.

Prabhakaran, S., Rey, M., Zagordi, O., et al. 2010. HIV-haplotype inference using a constraint-based dirichlet process mixture model. Machine Learning in Computational Biology (MLCB) NIPS Workshop 2010, 1–4.

Prosperi, M.C., Prosperi, L., Bruselles, A., et al. 2011. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics 12, 5.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38.

Rabiner, L.1989. A tutorial on hidden Markov models and selected applications in speech recognition (with erratum). Proceedings of the IEEE 77, 257–286.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78, 629–644.

Schultz, A.K., Zhang, M., Leitner, T., et al. 2006. A jumping profile hidden Markov model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics 7, 265.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Satist. 6, 461–464.

Spang, R., Rehmsmeier, M., and Stoye, J. 2002. A novel approach to remote homology detection: jumping alignments. J. Comput. Biol. 9, 747–60.

Tibshirani, R., Walther, G., and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B. 63, 411–423.

Vignuzzi, M., Stone, J., Arnold, J., et al. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439, 344–348.

Yamazaki, K., and Watanabe, S. 2005. Algebraic geometry and stochastic complexity of hidden Markov models. Neurocomputing 69, 62–84.

Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N. Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12, 119.

Zagordi, O., Geyrhofer, L., Roth, V., and Beerenwinkel, N. 2010. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J. Comput. Biol. 17, 417–28.

Zagordi, O., Klein, R., Däumer, M., and Beerenwinkel, N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 38, 7400–9.

Address correspondence to:
Prof. Niko Beerenwinkel
Department of Biosystems Science
and Engineering
ETH Zurich
Mattenstrasse 26
Basel 4058
Switzerland

E-mail: niko.beerenwinkel@bsse.ethz.ch