

# A bioinformatics pipeline for the analysis and interpretation of ultradeep sequence data

N Lance Hepler<sup>1,\*</sup>, Martin D Smith<sup>1</sup> Wayne Delport<sup>2</sup>, Jason A Young<sup>3</sup>, Art FY Poon<sup>4</sup>, Sergei L Kosakovsky Pond<sup>3</sup>

**1 Interdisciplinary Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, California, USA**

**2 Prognosys Biosciences, La Jolla, California, USA**

**3 Department of Medicine, University of California San Diego, La Jolla, California, USA**

**4 BC Centre for Excellence in HIV/AIDS, Vancouver, British Columbia, Canada**

**\* E-mail: Corresponding nhepler@ucsd.edu**

## Abstract

**Background:** Massively parallel sequencing technologies are increasingly adopted for the study of rapidly evolving and diverse pathogens, such as HIV-1. Custom bioinformatics tools are needed to address domain-specific challenges of viral sequence analysis and differentiate sequencing artifacts from biologically meaningful mutations. Large quantities of data require parallelized and efficient algorithms for rapid analysis.

**Results:** We have developed simple probabilistic methods to model ultradeep sequencing errors and implemented a web-based (<http://www.datamonkey.org>) and standalone analysis pipeline to facilitate common analyses of HIV-1 sequence data generated using the Roche 454 platform. Such analyses include robust read mapping and filtering, instrument error filtering, identification of drug-resistance associated and accessory mutations, diversity estimation, determination of whether a host is multiply infected, and the computation of simple measures of selective pressures on individual sites.

**Conclusions:** Our pipeline is designed to provide end-users with the tools to rapidly analyze sequences derived from the Roche 454 sequencing platform and is geared specifically for the analysis of HIV-1 and other rapidly evolving RNA viruses.

## Author Summary

## Introduction

Recent advances in DNA sequencing technology have facilitated the generation of massive amounts of data, the utility of which is entirely dependent on the analysis tools available to researchers. Ultradeep sequencing projects utilizing the Roche 454 Titanium FLX platform (the 454 platform) routinely generate tens or hundreds of thousands of reads from multiple genes and samples. In the context of HIV-1, these data have proven useful in studies of evolution of drug resistance [1, 2], response to host immune pressure [3–6], co-receptor usage [7, 8] and the dynamics of acute infection [9]. A major advantage of ultradeep sequencing platforms is their power to detect low-frequency, or minority, genetic variants. Population-based sequencing of plasma samples has been shown to be inefficient at identifying variants with prevalence less than 10% [10], whereas minority variants at frequencies below 1% have been recovered using ultradeep sequencing methods [11]. Given that circulating or archived minority variants can cause antiretroviral therapy failure [12–14], high-resolution sequence characterization of viral populations is desirable, particularly within the context of drug resistance and treatment. However, in doing so care must be taken in distinguishing true biological minority variants from artifactual variants arising due to PCR and sequencing errors. PCR misincorporation can introduce mutations at a rate of  $10^{-5}$  to  $10^{-4}$  [15], whereas base-calling and other 454 platform-specific errors are typically larger in regions containing repeats of three or more identical nucleotides (homopolymers) [1]. One approach to quantify ultradeep

sequencing error has been to sequence clonal samples [1], which suggested 0.0007 and 0.0044 mismatch errors per nucleotide sequenced in non-homopolymeric and homopolymeric regions, respectively. Since error rates are likely to be sample- and gene-specific, and since it is infeasible to obtain clonal control sequencing runs for all samples in a study, it is necessary to devise metrics for the detection of ultradeep sequence errors in a sample specific manner. To that end, we have developed simple methods to model ultradeep sequence errors and assembled an analysis pipeline to facilitate common analyses of HIV-1 sequence data generated using the 454 platform.

## Results

To illustrate the pipeline we present some results derived from a multiplexed 454 sequence data covering regions in *gag*, *rt* and *env* from a single plasma sample extraction. This sample was one of those recently analyzed for evidence of dual HIV infection [16].

### Alignment filtering and sequence diversity

Quality filtering with a PHRED score of 20 and minimum read length of 100 reduced the total number of reads from 19996 to 19377. However, since we split reads with internal regions of low quality, a total of 23824 high-scoring fragments were retained for subsequent analyses. Of these fragments, the majority mapped to the *env* reference sequence, followed by *gag* and *rt* (Table ??). In addition to these summary statistics, we produce plots of coverage and majority variants proportions along each gene region (Figure 1) to facilitate the rapid identification of regions of insufficient data quantity. Furthermore, the majority variant plots provide a means to identify regions of high diversity which may want to be targeted in subsequent ultradeep sequencing experiments. All plots are indexed according to the reference sequence (either HIV-1 or a custom upload), and are available for download from the server. Sliding windows of maximum sequence diversity, and neighbor-joining trees, further allow the user to identify regions of increased diversity, or in the case of HIV-1, identify poly-clonal infection (as in [16]). Since ultradeep sequence reads are typically short, and thus contain very little phylogenetic signal, we only estimate bootstrap support for neighbor-joining trees in the sliding window of maximum diversity. In most cases, however, there is little bootstrap support for any particular branching order in the tree.

### Sequence alignment

XXX TODO

### Estimation of 454 sequence error and mutation rate classes

Our binomial mixture model identified strong evidence for multiple mutation rate classes in all three genes with substantial AIC improvements in all cases (Table ??). **Assuming the lowest estimated mutation rate is the background mutation rate, we are able to identify sites that mutate at higher rates.** In some cases (as in site 234) the identification of true mutations is trivial, particularly when several variants occur at intermediate to high frequencies (Figure ??). However, we frequently need to distinguish minority variants from sequencing error, particularly when the former occur at frequencies less than 1%. For instance, sites 250 and 251 (Figure ??) are assigned to the background mutation rate class (in this case with rate = 0.002), whereas other sites (such as 262) with similar levels of diversity are assigned to a higher rate class. Given the coverage at a site, we can estimate the probability of observing  $m$  mutations assuming the background mutation rate (0.002 for *env*) using a binomial model (Figure ??). These results allow for the determination of whether allelic variants at a site are likely to be the result of sequencing error. This delineation is particularly important in the context of HIV minority

drug resistance and thus we utilize the same statistical tools to identify both sites, and variants at sites which can be separated from sequencing error. Results for all sites are presented as HTML pages and are also available as downloadable comma-separated value files for further processing.

## Discussion

Our pipeline is designed to provide end-users with the tools to rapidly analyze sequences derived from the 454 ultradeep sequencing platform and is geared specifically for the analysis of rapidly evolving populations, such as RNA viruses. We provide tools to rapidly filter reads based on reference sequences, to identify sequence variants within sliding windows, to identify allelic variants at sites, and to distinguish biologically meaningful variants from sequencing error artifacts. Estimates of site-specific purifying and diversifying selection are calculated, as are several statistics which facilitate the interpretation of sequence variation among ultradeep sequence reads. Finally, for HIV-1 sequence data we provide tools for the identification of polymorphisms at known drug-resistant sites. All these analyses tools are provided through an intuitive user-friendly interface at <http://www.datamonkey.org> and complement the array of other evolutionary biology tools available at our webserver. A standalone version of the pipeline is also provided, along with supporting documentation online.

## Methods

The HIV-1 analysis pipeline has been implemented as a HyPhy [17] Batch Language module, and on the Datamonkey webserver (<http://www.datamonkey.org>) [18, 19]. The pipeline comprises several phases, each of which interacts with a SQLite database (<http://www.sqlite.org>) backend for results storage and processing. Phases which can benefit from execution in parallel (e.g. read mapping) have been implemented to make use of distributed computing environments which use MPI.

### Quality filtering of reads

The analysis pipeline takes as input two files generated from a 454 sequencing experiment, the read file (.fna) in FASTA format (required), and the quality file (.qual) with site specific quality scores (q or PHRED) for each of the reads in the read file (optional). By default, reads that are at least 100 nucleotides long and have consecutive quality scores of 20 or greater are retained for subsequent analysis. These, and most other analysis parameters can be modified by the user. A quality score of 20 is interpreted as equivalent to one error in 100 bases (XXX) []. Errors in 454 pyrosequencing are frequently the result of incorrect determination of homopolymer lengths [1]. We make a special allowance for such regions to be excised by breaking a longer read with a poor quality score section in the middle into multiple fragments, each of which must meet the minimum length and quality score criteria.

### Sequence alignment

We use an iterative gene-specific alignment and filtering procedure. Firstly, we define a reference sequence, or set of reference sequences, which is used to filter “mappable” reads from the set of reads identified by the quality filtering step described above. A 454 sequencing run can include multiple amplicons of different gene regions without the need for multiplexing tags. Therefore, we provide a list of HIV-1 reference genes (HXB2, GenBank accession number K03455) which are initially used as a mapping reference. Alternatively, a custom reference sequence file can be uploaded. XXX REWRITE THIS ENTIRE SECTION Every read is translated into six protein sequences: one in each of the six (three forward and three reverse complement) reading frames. Encountered stop codons are mapped to the letter ‘X’ which is functionally equivalent to a missing base for the subsequence alignment phase. Next,

each translation is locally aligned [20] to the currently considered HXB2 gene, using an HIV-specific scoring matrix [21]. The reading frame with the highest alignment score is retained for subsequent steps. The reads are further filtered to select High Protein-Alignment Scoring (HPAS) reads, defined as reads whose *per-residue* alignment scores exceed the expected alignment score of a random sequence with sample-specific base composition by at least  $X$ -fold.  $X = 5$  yielded good filtering properties on synthetic data; this translates, roughly, into a  $\log_2 5$  informative bits per position (conceptually similar to BLAST bit scores). The consensus codon sequence of mapped HPAS reads is constructed and used as a sample-specific reference sequence from this point onwards. The remaining sequences, i.e. those not selected for the HPAS, are subsequently nucleotide aligned to the HPAS reference, and included if their alignment scores exceeds the median score of the distribution of HPAS reads. This nucleotide alignment step permits the correction of out-of-frame indels or homopolymer length errors specific to the 454 platform.

## Estimation of divergence

In order to identify distinct viral populations, e.g. in the context of dual infection [16], we estimate maximum sequence divergence in sliding windows with a default width and stride of 125 and 25 nucleotides, respectively. For each sliding window with minimum site coverage greater than a pre-defined value (default = 500), we estimate sequence divergence using the general time reversible model of nucleotide evolution [22], along the phylogeny inferred using the Neighbor Joining method [23]. Nucleotide diversity is determined as the maximum path length in the tree with branch lengths estimated by maximum likelihood. For this analysis, we require that a sequence variant occur at least  $N$  times (10 by default), or comprise at least  $P\%$  of the sample (1% by default), whichever is greater, to be considered a non-artifactual variant. We determine statistical support for internal nodes, and maximum sequence diversity, using bootstrap over alignment sites (100 replicates).

## Estimation of 454 sequence error and mutation rate classes

In order to estimate a sample-specific 454 sequence error rate, and thus a threshold for the identification of biologically meaningful minority polymorphisms at a site, we fit a binomial mixture model to site-specific mutation counts. A binomial distribution (albeit with an *a priori* error rate) has been suggested as a good PCR noise filtering model in the context of conservation genetics [24]. At site  $i$  from a 454 sample with coverage,  $c_i$ , the probability of observing exactly  $m_i$  mutations (identified as non-consensus amino acid residues at a site) assuming the binomial model is given by

$$L(D_i|r_1) = \binom{c_i}{m_i} r_1^{m_i} (1 - r_1)^{c_i - m_i}, \quad (1)$$

where the parameter,  $r_1$ —interpreted as the mutation rate at a site—is estimated by maximum likelihood as the proportion of observed counts. Assuming each site in an alignment is independent, the mean mutation rate can similarly be estimated for all sites as the product of site likelihoods using maximum likelihood, i.e.

$$L(D|r_1) = \prod_{i=1}^s L(D_i|r_1), \quad (2)$$

where  $s$  is the number of sites. We next consider a mixture model of  $K$  binomial distributions (each with its own rate  $r_j$ ). The likelihood of mutation data under this model is given by

$$L(D) = \prod_{i=1}^s \sum_{j=1}^K p_j L(D_i|r_j), \quad (3)$$

where  $p_j$ ,  $\sum_j p_j = 1$ , are the mixing proportions. The goodness-of-fit for each model is evaluated using Akaike Information Criterion (AIC). Beginning with  $K = 1$ , we increment the number of rate classes by one, until the model with  $K + 1$  rates is no longer preferred to the model with  $K$  rates by AIC. The mutation rates (and respective proportions) for each of the rate classes are presented, as are the improvements in model fit with the addition of rate classes.

Next, we *assume* that the class with the lowest mutation rate,  $r_e$ , represents sites subject to sequencing or instrument errors. This mutation rate can be used as a sample specific threshold for the identification of minority variants. More generally, we use an empirical Bayes procedure to obtain the posterior probability that site  $i$  is assigned to rate class  $j$  via

$$P(r_j|D_i) = \frac{L(D_i|r_j)P(r_j)}{\sum_{a=1}^k L(D_i|r_a)P(r_a)}, \quad (4)$$

where  $P(r_j) = p_j$  is the prior probability that site  $i$  belongs to rate class  $j$ . This approach is an improvement over *a priori* percentage or count-based thresholds, since it is dependent on both the depth (coverage) in an alignment, and on sample specific mutation patterns which are likely to vary between alignments (XXX) []. In addition to estimating the number of mutation rate classes supported by the data, we estimate the observed mutation rates at each of the sites in the alignment. Given a site with coverage,  $c$ , we estimate the effective mutation rate as the number of non-consensus amino acids,  $n$ , normalized by the coverage (i.e.  $n/c$ ).

At every site, we record the amino acid spectrum,  $A = (c_A, \dots, c_Y)$ , where the index iterates through the twenty amino acids and  $c_X$  denotes the count of amino acid,  $X$ , at the site. For visualization and data exploration purposes, we calculate the Shannon entropy at site  $s$ ,  $H(s)$ ,

$$H(s) = - \sum_{X, f_s(X) \neq 0} f_s(X) \log f_s(X),$$

where  $f_s(X)$  is the proportion of reads with residue  $X$  at site  $s$ , and the sum is taken over observed residues only.

## Sitewise estimation of diversifying and purifying selection

Typically estimates of diversifying and purifying selection are conditional on a known phylogeny (see [25, 26] for reviews). However, since ultradeep sequences from single 454 samples typically contain relatively short reads (200–300 bp) with low diversity ( $< 0.05$ ), the phylogenies estimated from the data have little statistical support. Therefore, we estimate selection at a site by considering the ratio of expected non-synonymous and synonymous mutations assuming neutral evolution, given the genetic code and observed codon frequencies, following the procedure outlined in detail in [27]. We also evaluate the ratio of the *observed* non-synonymous to synonymous substitutions at a site, by considering, for every read, the mean numbers of synonymous ( $s$ ) and non-synonymous substitutions ( $n$ ) along the shortest evolutionary path connecting it to every other read (i.e.  $c - 1$  possibilities, for  $c$  reads covering the site). This is equivalent to assuming a complete lack of phylogenetic information, i.e. any sequence is equally likely to be ancestral to any other sequence. Significance is assessed using the binomial distribution,

$$\binom{c}{s} p_s^s (1 - p_s)^{c-s},$$

as the probability of observing  $s$  or fewer synonymous substitutions out of  $c = n + s$  total substitutions given an expected proportion of synonymous substitution,  $p_s$ , under neutrality.

## Identification of Drug Resistant Variants

We screen for the occurrence of mutations in reverse transcriptase (*rt*), integrase (*int*) and protease (*pr*), that are known to confer resistance to antiretroviral agents ('drug resistance associated mutation' or DRAM sites). A current comprehensive list of such mutations is maintained at the Stanford HIV drug resistance database (<http://hivdb.stanford.edu>) [28]. For each DRAM site we rank its mutation rate with respect to all other sites (on a 0–100% scale), and calculate the median mutation rank of all non-DRAM sites. Given this median rank, we determine whether it is significantly elevated compared to that of non-drug resistant sites, using equivalently-sized random subsamples ( $n = 1000$ ) from sites not known to be directly associated with drug resistance. Such an elevation could indicate that selective forces are preferentially acting upon DRAM sites on average. We bin drug resistant sites into the estimated mutation rate classes to distinguish biologically and clinically relevant minority variants from those arising due to instrument error. Finally, we screen for known compensatory (accessory) mutations with the purpose of identifying whether these co-occur more frequently with known drug resistant variants than expected by chance. Such linkage between drug resistant and accessory mutations in the sample provides corroborating evidence that detected DRAMs are not artifactual.

## Acknowledgments

## References

1. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 17: 1195–201.
2. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, et al. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 35: e91.
3. Bimber BN, Burwitz BJ, O'Connor S, Detmer A, Gostick E, et al. (2009) Ultradeep pyrosequencing detects complex patterns of CD8+ T-lymphocyte escape in simian immunodeficiency virus-infected macaques. *J Virol* 83: 8247–53.
4. Hughes AL, O'Connor S, Dudley DM, Burwitz BJ, Bimber BN, et al. (2010) Dynamics of haplotype frequency change in a CD8+TL epitope of simian immunodeficiency virus. *Infect Genet Evol* 10: 555–60.
5. Poon AF, Swenson LC, Dong WW, Deng W, Kosakovsky Pond SL, et al. (2010) Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of hiv-1. *Mol Biol Evol* 27: 819–32.
6. Love TM, Thurston SW, Keefer MC, Dewhurst S, Lee HY (2010) Mathematical modeling of ultradeep sequencing data reveals that acute CD8+ T-lymphocyte responses exert strong selective pressure in simian immunodeficiency virus-infected macaques but still fail to clear founder epitope sequences. *J Virol* 84: 5802–14.
7. Archer J, Braverman MS, Taillon BE, Desany B, James I, et al. (2009) Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *Aids* 23: 1209–18.
8. Tsibris AM, Korber B, Arnaout R, Russ C, Lo CC, et al. (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE* 4: e5683.

9. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, et al. (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5.
10. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, et al. (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 43: 406–13.
11. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 17: 417–28.
12. Palmer S, Boltz V, Martinson N, Maldarelli F, Gray G, et al. (2006) Persistence of nevirapine-resistant HIV-1 in women after single-dose nevirapine therapy for prevention of maternal-to-fetal HIV-1 transmission. *Proc Natl Acad Sci U S A* 103: 7094–9.
13. Lecossier D, Shulman NS, Morand-Joubert L, Shafer RW, Joly V, et al. (2005) Detection of minority populations of HIV-1 expressing the K103N resistance mutation in patients failing nevirapine. *J Acquir Immune Defic Syndr* 38: 37–42.
14. Kapoor A, Jones M, Shafer RW, Rhee SY, Kazanjian P, et al. (2004) Sequencing-based detection of low-frequency human immunodeficiency virus type 1 drug-resistant mutants by an RNA/DNA heteroduplex generator-tracking assay. *J Virol* 78: 7112–23.
15. Kobayashi N, Tamura K, Aotsuka T (1999) PCR error and molecular population genetics. *Biochemical genetics* 37: 317–321.
16. Pacold M, Smith D, Little S, Cheng PM, Jordan P, et al. (2010) Comparison of methods to detect HIV dual infection. *AIDS research and human retroviruses* 26: 1291–1298.
17. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676–679.
18. Kosakovsky Pond SL, Frost SD (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21: 2531–3.
19. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26: 2455–7.
20. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48: 443–453.
21. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-specific probabilistic models of protein evolution. *PloS one* 2: e503.
22. Tavaré S (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences, *Amer Mathematical Society*, volume 17. pp. 57–86.
23. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4: 406–425.
24. Cummings S, McMullan M, Joyce D, van Oosterhout C (2010) Solutions for PCR, cloning and sequencing errors in population genetic analysis. *Conservation Genetics* 11: 1095–1097.
25. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. *Briefings in bioinformatics* 10: 97–109.

26. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* 26: 255–71.
27. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution* 22: 1208–1222.
28. Shafer RW (2006) Rationale and uses of a public HIV drug-resistance database. *The Journal of infectious diseases* 194 Suppl 1: S51–S58.

## Figure Legends

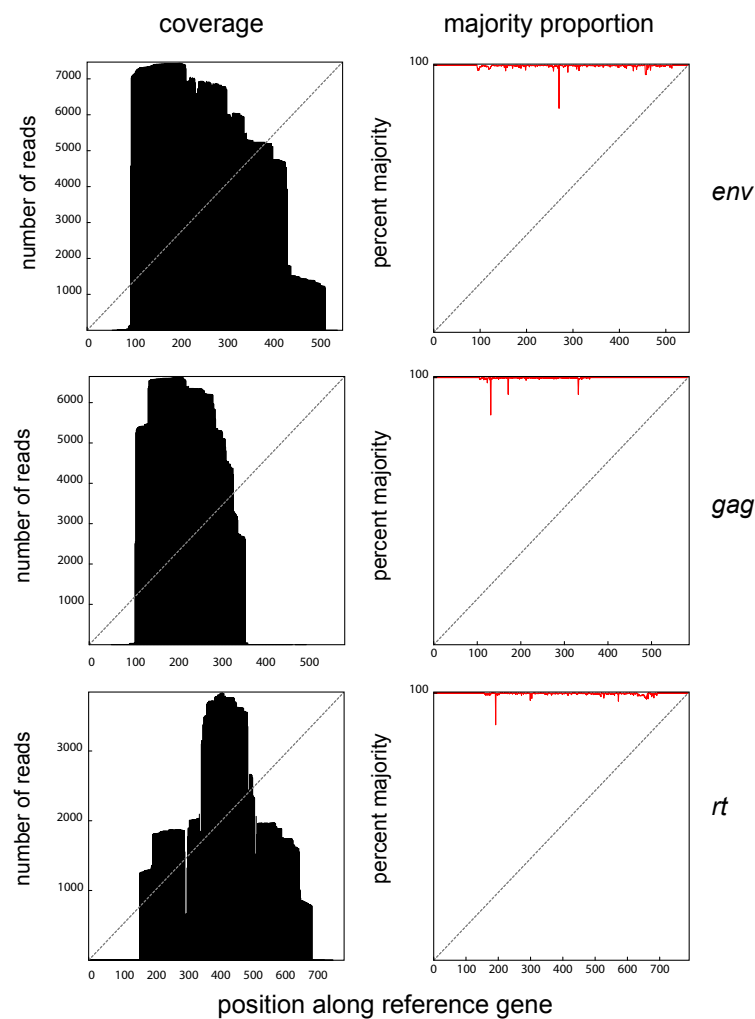
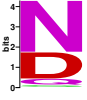

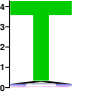

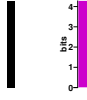



Figure 1. Site-specific coverage and majority variants obtained in a single 454 run containing *env*, *gag*, and *rt* amplicons.



Site	234	238	244	250	251	262
						
Minority (aa)	1 (F,V,E) 2 (K,G) 21 (S)* 67 (Q)* 271 (D)*	1 (I,A) 2 (R) 4 (E) 6 (T) 10 (L)* 12 (S)* 22 (D)* 23 (Q)* 266 (M)*	1 (K,Y) 2 (E) 4 (S) 12 (I)* 65 (P)* 98 (Q)* 127 (H)* 289 (A)*	1 (T) 2 (R) 8 (E) 13 (W)*	1 (M) 3 (L) 4 (F) 6 (T) 11 (V)*	1 (K) 3 (Y) 10 (S) 20 (D)*
Majority (aa)	523 (N)	6227 (P)	6990 (T)	7317 (G)	7331 (I)	7372 (N)
Coverage	889	6574	7289	7341	7356	7406
Rate	0.434	0.054	0.088	0.002	0.002	0.007
P	1.0	1.0	1.0	0.99	0.99	0.98

**Figure 2. Assignment of *env* sites to rate classes based on a binomial mixture model.**

Examples of sites with intermediate and low-frequency variants are shown to demonstrate the ability of the model to distinguish mutations at a site from a background (or 454 error) mutation rate. Shown are the amino acid profiles, the coverage, the mutation rate of the class to which the site is assigned, and the posterior probability,  $P$ , that the site belongs to the assigned rate class. \* Indicates significance ( $P \leq 0.05$ ) for a test of whether the  $n$  observed mutations are expected to occur given the estimated background mutation rate (0.002) and the coverage at a site.

## Tables

**Table 1. Summary read statistics from a 454 sample containing *env*, *gag*, and *rt* sequences (sd = standard deviation)**

gene	reads	mean read length	sd read length	mean coverage at a site	sd coverage
<i>env</i>	8620	266.37	92.65	4182.46	2992.71
<i>gag</i>	6728	214.48	40.26	2466.65	2921.27
<i>rt</i>	6683	179.64	66.70	1521.62	1304.18

**Table 2. Mutation rate classes estimated using the binomial mixture model**

gene	number of rate classes	mutation rates range	proportion background	$AIC_i$
<i>env</i>	8	0.002–0.434	0.55	28032.36
<i>gag</i>	4	0.002–0.175	0.84	13756.25
<i>rt</i>	5	0.002–0.126	0.46	5938.45

$AIC_i$  is the improvement in fit over a single-rate model; proportion background is the proportion of sites which are assigned to the background, or the smallest estimated mutation rate.