Experimental Analysis of Sources of Error in Evolutionary Studies Based on Roche/454 Pyrosequencing of Viral Genomes

Ericka A. Becker[1], Charles M. Burns[2], Enrique J. León[1], Saravanan Rajabojan[3], Robert Friedman[3], Thomas C. Friedrich[1,4], Shelby L. O'Connor[2], and Austin L. Hughes[3,*]

[1]Wisconsin National Primate Research Center, University of Wisconsin

[2]Department of Pathology and Laboratory Medicine, University of Wisconsin

[3]Department of Biological Sciences, University of South Carolina

[4]Department of Pathobiological Sciences, University of Wisconsin

*Corresponding author: E-mail: austin@biol.sc.edu.

Accepted: 13 March 2012

Abstract

Factors affecting the reliability of Roche/454 pyrosequencing for analyzing sequence polymorphism in within-host viral populations were assessed by two experiments: 1) sequencing four clonal simian immunodeficiency virus (SIV) stocks and 2) sequencing mixtures in different proportions of two SIV strains with known fixed nucleotide differences. Observed nucleotide diversity and frequency of undetermined nucleotides were increased at sites in homopolymer runs of four or more identical nucleotides, particularly at AT sites. However, in the mixed-strain experiments, the effects on estimated nucleotide diversity of such errors were small in comparison to known strain differences. The results suggest that biologically meaningful variants present at a frequency of around 10% and possibly much lower are easily distinguished from artifacts of the sequencing process. Analysis of the clonal stocks revealed numerous rare variants that showed the signature of purifying selection and that elimination of variants at frequencies of less than 1% reduced estimates of nucleotide diversity by about an order of magnitude. Thus, using a 1% frequency cutoff for accepting a variant as real represents a conservative standard, which may be useful in studies that are focused on the discovery of specific mutations (such as those conferring immune escape or drug resistance). On the other hand, if the goal is to estimate nucleotide diversity, an optimal strategy might be to include all observed variants (even those at less than 1% frequency), while masking out homopolymer runs of four or more nucleotides.

Key words: pyrosequencing, natural selection, simian immunodeficiency virus, homopolymer.

Introduction

Pyrosequencing by technologies such as 454 has the potential to contribute to numerous areas of biology (Langaee and Ronaghi 2005; Margulies et al. 2005; Ahmadian et al. 2006; Bushman et al. 2008). Among these is the analysis of the population genetics of viral populations within hosts over the course of infection (Bimber et al. 2009, 2010; Rozera et al. 2009; Tsibris et al. 2009; Hedskog et al. 2010; Hughes et al. 2010; Poon et al. 2010; Wang et al. 2010). This type of analysis may provide crucial information for understanding the pathogenesis of certain RNA viruses, such as human immunodeficiency virus-1 (HIV-1) and hepatitis C virus (HCV). Not only do HIV-1 and HCV have high mutation rates, as is typical of RNA viruses, but they are successful in evading the ordinary antiviral immune defenses, leading to persistent infection in the vast majority of HIV-1 infections and probably in a majority of HCV infections as well.

There is evidence that mutations conferring escape from recognition by host CD8+ T-lymphocytes (CD8+TL) are selectively favored during infection with immunodeficiency viruses and HCV, and such escape from immune recognition has been hypothesized to play a role in the persistence of these viruses (Allen et al. 2000; Erickson et al. 2001; O'Connor et al. 2004; Guglietta et al. 2005). In addition, mutations in HIV-1 that confer resistance to antiretroviral treatment are an ongoing public health concern (Martinez-Picardo et al. 2000; Truong et al. 2006). By providing data on the sequence variants occurring within a host, their frequencies, and the changes in those frequencies over the course of infection, pyrosequencing can help identify population processes (such as natural selection and genetic drift) that play key roles in determining the fate of immune escape and drug resistance mutations and thus provide an enhanced understanding of the factors responsible for viral persistence (Hughes et al. 2010).

As with any technology, pyrosequencing with Roche/454 technologies has certain known sources of error. One such source is the tendency toward incorrect base calls in homopolymer regions; that is regions where the same base is repeated multiple times (Margulies et al. 2005). We studied the effect of this and related sources of error on estimating within-host virus population parameters, such as nucleotide diversity, by analyzing Roche/454 pyrosequencing data from clonal and mixed strains of simian immunodeficiency virus (SIV). By quantitatively examining the sources and magnitude of error, the results provide background information that will aid researchers in interpreting data derived from this important technology.

Although numerous studies have published estimates of error rate in pyrosequencing (Wang et al. 2007; Wiseman et al. 2009), these studies focus mainly on the artifacts arising when sequencing a homogeneous sequence sample. Particularly in the case of RNA viruses, within-host samples of virus populations are far from homogeneous. This nonhomogeneity may affect pyrosequencing error rates in unknown ways. Moreover, studies on within-host virus evolution pose the analytic challenge of assessing the extent to which artifacts mask genuine patterns of nucleotide substitution.

In pyrosequencing studies of within-host viral populations, the impact of artifacts may differ depending on the type of question being addressed. For example, studies that seek to discover novel CD8+TL escape or drug resistance mutations may be concerned with whether certain low-frequency variants represent artifacts or real mutations (Wang et al. 2007). By contrast, certain other studies may be less interested in individual sequencing artifacts than in the possible impact of multiple artifacts on estimates of genome-wide nucleotide diversity. This might be true, for example, in studies that test hypotheses regarding the relationship between levels of virus sequence diversity and clinical outcome. Likewise, in evolutionary studies that estimate patterns of synonymous and nonsynonymous nucleotide substitution in an effort to understand patterns of natural selection on protein-coding regions, it is of interest to know how these estimates are affected by sequencing artifacts. In the present study, we focus on a known source of error—that caused by homopolymer runs—in order to examine the impact of sequencing artifacts on the diverse array of questions regarding within-host viral diversity that can be addressed by pyrosequencing.

Materials and Methods

Preparation of Viral Stocks

The SIVmac239 and the mutant SIV stock (termed m3KO) were prepared similarly. Separate plasmids containing hemigenomes of SIVmac239 (accession M33262) were originally obtained from the NIH AIDS Reference Reagent Program (deposited by R. Desrosiers). The m3KO virus contained 24 single nucleotide differences relative to SIVmac239 (manuscript in preparation). Plasmid hemigenomes of m3KO were prepared by custom gene synthesis (Genscript), and they were designed with the same restriction sites as the SIVmac239 plasmids. Each plasmid was digested with SphI and then religated to make a full-length provirus. The ligation mixture was transfected into Vero cells and infectious SIV was rescued. Viral stocks were prepared as previously described (Valentine et al. 2009). To prepare the four SIVmac239 viral stocks used in the diversity analysis, rescued virus was expanded on rhesus macaque peripheral blood mononuclear cells that had been activated by pulsing with concanavalin A. To prepare the SIVmac239 and m3KO viruses for the mixing experiment, transfected Vero cells were cocultured with CEMx174 cells for 48 h. The transduced CEMx174 cells were grown for 10–12 days, and each virus was harvested.

Viral Sequencing

Roche/454 pyrosequencing of viruses was performed, essentially, as previously described (Bimber et al. 2010; O'Connor et al. 2012). For summary statistics on sequencing results, see supplementary table S1 (Supplementary Material online). The goal of this sequencing strategy is to provide numerous sequence reads that span entire coding sequence (CDS) of the viral genome. Because each nucleotide site in the region covered is represented numerous times in the resulting reads, this strategy provides us an in-depth picture of nucleotide diversity at individual sites (Bimber et al. 2010). No de novo sequence assembly or consensus sequence was generated because the purpose of this approach is to examine within-host viral population diversity. Viral RNA was isolated using the Qiagen MinElute viral RNA isolation kit (Qiagen). Four overlapping amplicons spanning the entire SIV genome were prepared from viral RNA using the Superscript III One-Step RT-PCR System with Platinum Taq High Fidelity (Invitrogen). Polymerase chain reaction (PCR) products were purified and then libraries were created using the Nextera DNA Sample Prep kit (Epicentre) and labeled with Multiplex Identifier (MID) Tags. Libraries were pyrosequenced with a Roche/454 GS Jr instrument and Titanium shotgun chemistry, according to the manufacturer's protocols (454 Life Sciences).

Calculation of Proportion Variant and Undetermined Nucleotides

Nucleotide sequence alignments were performed with a suite of tools available at a local installation of the Galaxy software (Blankenberg et al. 2010; Goecks et al. 2010). Sequence reads were base called with Roche base caller version 2.5p1, and the sff files were converted to FASTQ files. Low-quality sequences at the 3′ ends were removed during conversion. Adaptor, MID, and transposon sequences were then trimmed from FASTQ files. We masked low quality bases (quality < 18) and reads were then mapped to SIVmac239 (Accession # M33262) using LASTZ at a 90% identity threshold (Harris 2007). This level of identity is sufficiently lax to allow for viral mutation but sufficiently strict to preclude aligning of contaminants of cellular origin. The percent variation at each nucleotide position was calculated with SAMtools, excluding nucleotide sites that were masked (Li et al. 2009). The percent of deletion/insertion polymorphisms (Dip percent) and undetermined nucleotides (proportion N) were also calculated. Data was imported into a local installation of LabKey software (Nelson et al. 2011). These data were filtered to create a report for each viral genome that included the number of A, C, T, and G and the proportion variant, Dip, and N nucleotides at each position.

Nucleotide Diversity

At each individual nucleotide site, the nucleotide diversity (π) was estimated by the number of pairwise nucleotide differences (n_d) divided by the total number of pairwise comparisons. In this computation, we included only nucleotides for which a base assignment (A, C, T, or G) was made. If n is the number of such nucleotides for a given site, then for that site $\pi = 2n_d/(n^2 - n)$. The π values were averaged across all sites in order to obtain a mean π value for the entire region sequenced.

In the case of CDSs, we computed π separately for synonymous and nonsynonymous sites. The sum of the individual π values for all synonymous sites was then divided by the total number of synonymous sites (Nei and Gojobori 1986) in a given open reading frame in the reference sequence in order to obtain the synonymous nucleotide diversity (π_S) for that reading frame. Similarly, the sum of the individual π values for all nonsynonymous sites was then divided by the total number of nonsynonymous sites in a given open reading frame in the reference sequence in order to obtain the nonsynonymous nucleotide diversity (π_N) for that reading frame.

The % variant at a given site was defined as the percent of nucleotides (excluding undetermined nucleotides) differing from the reference SIVmac239 sequence. In certain analyses, we excluded sites at which the % variant was less that 1%; in other words, for the latter analyses, such sites were treated as invariant sites. The 1% cutoff has been used in previous analyses of Roche/454 pyrosequencing data as an ad hoc rule of thumb for including variants likely to be biologically meaningful (e.g., Bimber et al. 2010). By comparing the results of analyses that included variants at less than 1% frequency with those of analyses excluding those variants, we assessed the impact of this rule of thumb on inferences regarding patterns of sequence polymorphism.

Mixed Strains

We decided to mix two virus strains of known sequence at various ratios in order to determine whether our sequencing and analysis approach could reliably detect polymorphisms at an expected frequency throughout the genome. Depending on the strain combination, we analyzed 6,073–6,087 sites. These sites were derived from positions1290–5659 and positions 8488–10205 of the reference sequence; these were regions where we had a sufficient amount of input material for the sequencing reaction, such that template resampling was not a concern. There were 17 known fixed nucleotide differences between the two strains in the sites analyzed. The number of viral genomes per milliliter of each stock was quantified by quantitative RT-PCR. The strains were mixed prior to the isolation of vRNA in the following ratios: 1:19, 1:4, 1:2, and 1:1

General Linear Models

General linear models analysis of variance (ANOVA) was conducted in Minitab version 15.0 (http://www.minitab.com). In these analyses, the individual nucleotide site was the unit of analysis. In analyses of the viral stocks, sites were classified by two predictor (independent) variables: 1) Run Length, the length of the run of identical nucleotides (homopolymer run) in which the site was located and 2) Reference bp, the base pair (AT) or (GC) at that site in the reference sequence. In referring to homopolymer runs, we used the shorthand H2 to indicate a homopolymer run of two nucleotides, H3 to indicate a homopolymer run of three nucleotides, and so forth. H1 indicated sites not in homopolymer runs. The dependent variables in these analyses were mean π for the four viral stocks and the mean proportion of undetermined nucleotides (Prop. N) for the four viral stocks.

In the data from the mixed-strain experiment, general linear models analysis of covariance (ANCOVA) was used to test for differences between sites with known fixed differences between the strains, controlling for the effect of homopolymer runs. The variable Strain Diff. categorized sites depending on whether or not the site was 1 of the 17 sites with known differences between the strains. Because there were only 17 sites with fixed strain differences, such sites were not available for all categories of Run Length (H1–H6). Therefore, in these analyses, Run Length was included as a covariate rather than as a classificatory factor.

Results

Nucleotide Diversity in Viral Stocks

Four SIVmac239 stocks were pyrosequenced, surveying 8,877 sites spanning the entire SIV genome in each of the four stock samples. Across all four stocks, there were a total of 7,198 sites at which polymorphism was detectable in one or more of the stock samples. The number of variable sites per stock ranged from 2,464 to 4,883. However, the vast majority of variants were at frequencies less than 1%. When only variants at 1% frequency or greater were included, there were 367 sites at which polymorphism was detectable in one or more of the stock samples, and the number of variable sites per stock ranged from 96 to 150. In the reference sequence for the region analyzed, there were 1,288 H2 runs, 343 H3 runs, 131 H4 runs, 32 H5 runs, and 8 H6 runs. Of H2, 791 (61.4%) involved A or T; of H3, 224 (65.3%) involved A or T; of H4, 94 (71.5%) involved A or T; of H5, 24 (75.0%) involved A or T; and of H6, 7 (87.5%) involved A or T.

A factorial ANOVA was used to assess factors contributing to nucleotide diversity (π) at individual sites. These analyses were conducted separately for each of the four stocks and then on mean values for the four stocks. Because the results for the individual stocks were similar (data not shown), we report here only the results for the analysis of mean π for all four stocks. The analysis showed a significant effect on mean π of Run Length ($F_{5,8865} = 57.87$, $P < 0.001$). There was little difference in mean π among H1 (0.00240 ± 0.00005), H2 (0.00226 ± 0.00005), and H3 (0.00241 ± 0.0010). But mean π was markedly elevated in H4 (0.00384 ± 0.00036), H5 (0.01085 ± 0.00206), and H6 (0.01743 ± 0.00590). There was also a significant main effect of Reference bp ($F_{1,8865} = 63.48$, $P < 0.001$), explained mainly by the higher mean π at AT sites than at GC sites (fig. 1*A*).

There was a highly significant interaction between Run Length and Reference bp ($F_{5,8865} = 11.02$, $P < 0.001$; fig. 1*A*). The mean π value at AT sites in H4, H5, and H6 was particularly elevated in comparison to GC sites (fig 1*A*). In addition, mean π at GC sites in H6 was actually lower than that in H5 (fig. 1*A*). However, there was only one H6 run of GC sites, so that chance factors may have contributed to the low observed mean π value (fig. 1*A*).

We conducted the same analysis excluding sites less than 1% variant, and the results were essentially the same. Again there were significant main effects on mean π of both Run Length ($F_{5,8865} = 52.05$, $P < 0.001$) and Reference bp ($F_{1,8865} = 72.98$, $P < 0.001$). Likewise, there was a significant interaction between Run Length and Reference bp ($F_{5,8865} = 14.33$, $P < 0.001$; fig. 1*B*). As when all variant sites were included, this interaction was explained by the unusually high mean π at AT sites in H4, H5, and H6 (fig. 1*B*).

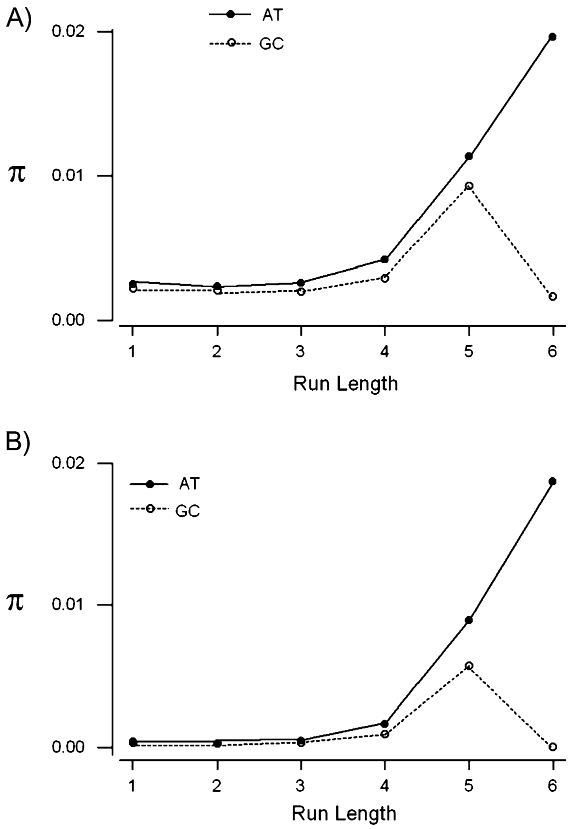

Fig. 1.—Mean nucleotide diversity (π) of four SIV stocks at sites categorized by length of homopolymer runs (Run Length) and the reference sequence base pair (AT or GC; Reference bp): (*A*) including all variable sites and (*B*) excluding sites with % variant < 1%. The figure illustrates the significant interactions in factorial ANOVA between the variables Run Length and Reference bp.

Proportion of Undetermined Nucleotides

A factorial ANOVA, using Run Length and Reference bp as factors, was applied to the mean proportion of undetermined nucleotides (Prop. N) at all variable sites. There were highly significant main effects of Run Length ($F_{5,7186} = 1893.82$, $P < 0.001$) and of Reference bp ($F_{1,7186} = 9.55$, $P = 0.002$). The effect of Run Length was explained by a steady increase in mean Prop. N as a function of homopolymer run length (fig. 2). The significant effect of Reference bp was due to consistently higher mean Prop. N at AT sites than at GC sites, across all lengths of homopolymer runs (fig. 2). There was also a highly significant interaction between Run Length and Reference bp ($F_{5,7186} = 3.90$, $P = 0.002$; fig. 2). This interaction was explained by the fact that the difference in mean Prop. N between AT sites and GC sites increased as the length of the homopolymer run increased (fig. 2).

Synonymous and Nonsynonymous Substitution

Because artifactual changes are expected to occur at random with respect to the reading frame of genes, they should not be expected to show the effect of purifying selection,

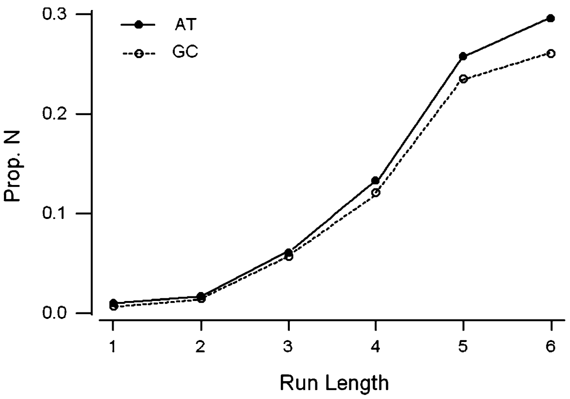

Fig. 2.—Mean proportion undetermined nucleotides (Prop. N) of four SIV stocks at sites categorized by length of homopolymer runs (Run Length) and the reference sequence nucleotide. The figure illustrates the significant interaction in factorial ANOVA between the variables Run Length and Reference bp.

which reduces nucleotide diversity at nonsynonymous sites by eliminating deleterious nonsynonymous mutations. Therefore, we tested whether observed patterns of nucleotide diversity showed an effect of purifying selection by estimating mean synonymous (π_S) and nonsynonymous (π_N) nucleotide diversity in the nine SIV protein-coding reading frames for the four viral stocks. We conducted these analyses in two ways: 1) including all observed variant sites and 2) including only variant sites at which the % variant was at least 1% (table 1). In both cases, the analyses were also conducted excluding observed variants at sites in H $\geq$ 4 (i.e., homopolymer runs of four or more; table 1). We categorized sites in this way because the numbers of sites were small if H4, H5, and H6 were analyzed separately and because both π and Prop. N seemed to increase markedly in H4 compared with H2 and H3 (figs. 1 and 2).

In each case, mean π_S was significantly greater than mean π_N, a pattern indicative of purifying selection (table 1). Excluding substitutions in H $\geq$ 4 led to a significantly reduced π_N when all variable sites were included and to a significant reduction in both π_S and π_N values when only sites with percent variant $\geq$ 1% were included (table 1). Both in the case of all sites and in the case of sites with percent variant $\geq$ 1%, excluding substitutions in H $\geq$ 4 led to reduced π_N:π_S ratios (table 1). Both π_S and π_N values were five to nine times greater when all variants were included than when we included only variants with % variant $\geq$ 1% (table 1). The lowest π_N:π_S ratio (0.512) was seen when both variants at less than 1% frequency and variants in H $\geq$ 4 were excluded (table 1).

Table 1
Mean Synonymous (π_S) and Nonsynonymous (π_N) Nucleotide Diversity in Clonal SIVmac239 Stocks

	π_S ± SE	π_N ± SE	π_N:π_S
All observed variants			
All substitutions	0.00366 ± 0.00013	0.00236 ± 0.00007***	0.645
Excluding H $\geq$ 4^a	0.00349 ± 0.00013	0.00202 ± 0.00005***,†	0.579
% Variant $\geq$ 1% only			
All substitutions	0.00063 ± 0.00006	0.00046 ± 0.00002**	0.730
Excluding H $\geq$ 4^a	0.00041 ± 0.00006†	0.00021 ± 0.00002**,†	0.512

Note.—SE, standard error.

[a] Homopolymer runs of four or more nucleotides.

Paired *t*-test of the hypothesis that $\pi_S = \pi_N$: $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Paired *t*-test of the hypothesis that π_S or π_N for H $\geq$ 4 equals the corresponding value for all substitutions: $^\dagger P < 0.001$.

Mixed Strains

In mixed-strain experiments, the expected π value for each mixture was determined based on the number of determined nucleotides at each of the sites and the expected proportions of each strain in the mixture. The expected mean π values for the 17 sites in each of the mixtures, based on the expected proportions 1:19, 1:4, 1:2, and 1:1, were, respectively, 0.0952, 0.3206, 0.4452, and 0.5012. The observed mean π values for the 17 sites in each of the mixtures were, respectively, 0.1846, 0.3523, 0.4946, and 0.4299. Thus, the 1:4 mixture yielded the observed mean π closest to the expected value, whereas the 1:2 mixture was the next closest to the expected value.

For each mixture, ANCOVA was used to test for an effect on π of known strain differences (variable Strain Diff.), with Run Length as a covariate. In every case, there was a highly significant main effect of the covariate Run Length, a highly significant main effect of Strain Diff., and a highly significant interaction between Run Length and Strain Diff. ($P < 0.001$ in every case). In all four mixtures, mean π was much higher at the sites with known strain differences than at other sites (fig. 3). The significant Run Length-by-Strain Diff. interactions imply that the slope of the linear relationship between π and Run Length differed between sites with a strain difference and sites without a strain difference (fig. 3). In the case of the 1:2, 1:4, and 1:19 mixtures, the slope was more strongly positive in the case of sites with a strain difference than in other sites (fig. 3*B–D*). By contrast, in the case of the 1:1 mixture, the slope in the case of sites with a strain difference was negative (fig. 3*A*).

There was just one site with a known strain difference that fell on a homopolymer of four or more nucleotides; this site was in a H6 run of A's (fig. 3). In order to test whether this point was influential in the ANCOVA results, we conducted each analysis excluding that point. In every case, the effects of Strain Diff. and Run Length were still significant ($P < 0.001$ in each case). On the other hand, the Run Length-by-Strain Diff. interactions were not significant in every case when the point in question was excluded. In the 1:1 and 1:4 mixtures, the Run Length-by-Strain Diff. interactions

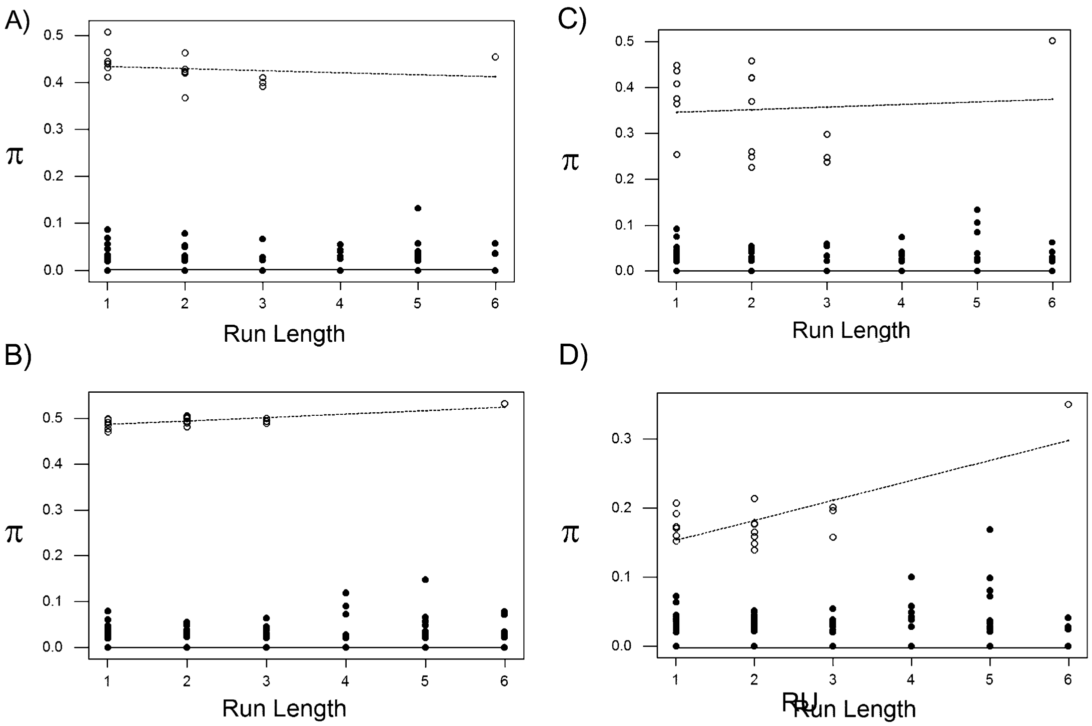

Fig. 3.—Nucleotide diversity (π) at individual variable sites plotted against homopolymer run length in four strain combinations: (*A*) 1:1, (*B*) 1:2, (*C*) 1:4, and (*D*) 1:19. In each case, open circles indicate sites (N = 17) with known strain differences and closed circles indicate all other sites. Separate linear regression lines are drawn for the sites with known strain differences (dotted lines) and all other sites (solid lines).

were still significant at $P < 0.001$. In 1:2 mixture, the Run Length-by-Strain Diff. interaction was significant at $P = 0.024$. Only in the 1:19 mixture, was the Run Length-by-Strain Diff. interaction not significant. The latter result can be explained by the fact that, in the case of the 1:19 mixture, the remaining 16 sites with known strain differences showed no obvious linear trend as a function of Run Length (fig. 3*D*).

The percent of the total variance in π accounted for by each main effect and the interaction was estimated by expressing the sequential sum of squares as a percentage of the total sum of squares. In every case, by far the highest percentage of the overall variance was accounted for by Strain Diff. In the 1:19 mixture, the effect of Strain Diff. accounted for 76.9% of the overall variance in π (fig. 4). In the 1:4 mixture, the effect of Strain Diff. accounted for 86.4% of the overall variance in π (fig. 4). In the 1:2 and 1:1 mixtures, the main effect of Strain Diff. accounted for, respectively, 96.5% and 96.6% of the overall variance in π (fig. 4). By contrast, the covariate Run Length accounted for, respectively, 0.38%, 0.10%, 0.11%, and 0.07% of the total variance. Thus, in every mixture, the known strain differences produced a much stronger signal than any other factor, including the effect of homopolymer runs.

For the four mixtures, the percentage of the variance accounted for by Strain Diff. was positively correlated with the observed mean π at the 17 sites with fixed differences (adjusted $r^2 = 91.8\%$, $P = 0.018$; fig. 4). The regression equation ($Y = 63.7667 + 69.20300X$; fig. 4) was used to predict the percentage of the variance in π expected to be accounted for by Strain Diff., given a value of mean π at the 17 sites. With mean π = 0.05, Strain Diff. was predicted

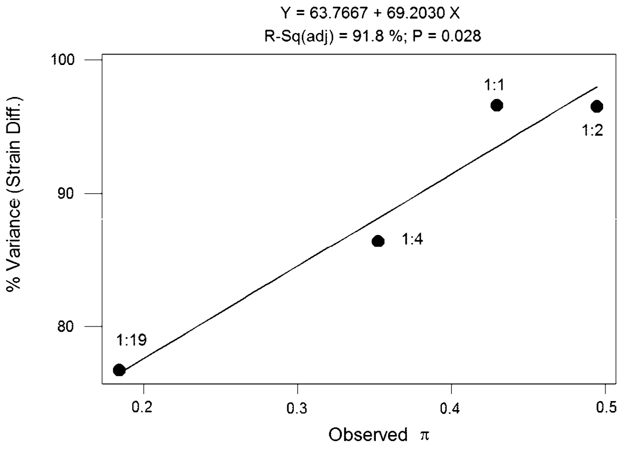

Fig. 4.—Regression of the percentage of the overall variance in π that is accounted for by fixed strain differences against the observed mean π at sites with known strain differences. The expected strain ratio is indicated for each point.

to account for 67.2% of the variance in π and with mean $\pi = 0.01$, Strain Diff. was predicted to account for 64.5% of the variance in π. These estimated values suggest that, under the conditions of these mixed-strain experiments, even if one of the two strains were present in very small proportions, over half of the variance in mean π would still be accounted for by strain differences.

As a test of the predictive value of the regression approach, we conducted four separate regression analyses, leaving out one of the mixtures in each case. Then, we used the regression model to predict the percent variance accounted for by Strain Diff. for the mixture that was left out. The predicted percent variance for the mixture that was left out was never more that 5% different from the observed percent variance. These results suggest that the predictive value of the regression method was good even though the number of observations was small.

Discussion

In this study, we used a statistical analysis of Roche/454 pyrosequencing data of laboratory stocks of SIVmac239 in order to improve our definition of biologically meaningful polymorphisms. We found both higher frequencies of observed substitution and higher proportions of undetermined nucleotides (N's) at sites in runs of four or more identical nucleotides. This is consistent with previous studies indicating that long homopolymer tracts contain regions of low confidence (Kircher and Kelso 2010). However, no detectable effect was seen in homopolymer runs of just two or three bases in length. The effect of runs of four of more identical nucleotides was most marked in the case of runs of A's or T's. A similar result has been reported in pyrosequencing of plant organelle genomes (Moore et al. 2006). A possible explanation relates to the fact that, when the synthetic A is incorporated during the sequencing process, the signal to noise ratio detected by the camera is not as high as for the other nucleotides (Ahmadian et al. 2006; Kajiyama et al. 2011).

On the other hand, sequencing of mixtures of two SIV strains with known nucleotide differences showed that the effects of apparent artifacts on the overall pattern of sequence diversity were small compared with that of known differences. In four different experiments, with observed nucleotide diversity (π) at sites with known strain differences ranging from about 0.18 to 0.49, by far the greatest proportion of variance among sites was accounted for by known strain differences. Even though a significant effect of occurrence in homopolymer runs was detected, the latter effect accounted for a much smaller percentage of the variance in π across sites than did the known strain differences. These results imply that biological variants with nucleotide diversity of 18% (the lowest observed π value in the different mixtures) or more can easily be detected and are readily distinguished from artifacts. We were not able to address directly whether this would be true of rarer variants, and future experiments will be needed to test the effect of less common variants. However, a regression approach based on our current data suggested that biologically real variants with nucleotide diversities as low as 5% or even 1% might be expected to provide a substantially greater signal than artifacts, even those due to homopolymer runs.

Because experimentalists may be more used to thinking in terms of percent variant than nucleotide diversity, it may be useful to mention how these quantities relate in data similar to those reported here. At the average site in the mixed-strain experiments, there were about 840 determined nucleotides (840× coverage). Given that number of nucleotides, a nucleotide diversity of 18% corresponds to a percent variant of about 10.1%. Given the same number of nucleotides, a nucleotide diversity of 1% corresponds to a percent variant of about 0.6%. Thus, our results suggest that a variant at 10% frequency is almost certain to be real and suggest that variants at 1% frequency have a good chance of being real. These guidelines would not apply if coverage is very low (<100) because in that case, stochastic error might yield artifacts with frequencies of as much as 1–10%. However, if coverage is several hundred folds, as in the present study, the above guidelines seem reasonable.

In both of the experiments, numerous low-frequency variants were detected by sequencing, and it could not be determined with certainty in the case of a given variant whether it represented a genuine mutation or a sequencing artifact. However, in the case of the four unmixed viral stocks, synonymous nucleotide diversity (π_S) was significantly greater than nonsynonymous nucleotide diversity (π_N), a pattern indicative of purifying selection (table 1). These results support the hypothesis that many observed low-frequency variants represent real mutations rather than sequencing artifacts. The pattern of $\pi_S > \pi_N$ would not be expected if all the variation observed represented sequencing artifacts because presumably, such artifacts would occur at random with respect to reading frames. On the other hand, $\pi_S > \pi_N$ is expected in the case of real polymorphisms because most nonsynonymous mutations are deleterious and are eliminated by purifying selection (Hughes 1999). Likewise, the mixed-strain populations showed numerous variants that were not in homopolymer runs but were at frequencies much lower than the known strain differences (fig. 3). The latter also may include many that represent real mutations, even though none approached the known strain differences in frequency.

Computer simulation of pyrosequencing data is an important tool for testing how bioinformatics applications handle sequence data (Lysholm et al. 2011). At the same time, statistical models have been developed to filter noise from real sequence differences (Quince et al. 2011) and to estimate allele frequencies in the presence of sequencing error (Lynch 2009). Empirical studies of pyrosequencing samples from biological organisms, such as the present study, can help improve both simulation packages and statistical models by providing insight into how sequence properties affect the observed results in real-life situations. They also suggest strategies that empirical researchers may use to minimize the impact of error, given the type of biological question in which they are interested.

Our results imply that the impact of pyrosequencing artifacts differs depending on the type of question being addressed. Studies that are interested in discovering biologically important new variants, such as CD8+TL escape or drug resistance mutations, may want to use a relatively conservative cutoff, such as 1%. In the case of mutations in homopolymer runs of four or more nucleotides, particularly those involving AT base pairs, an even higher cutoff may be desirable. Note that in the case of putatively selectively favored mutations, other data may help to confirm the hypothesis that a given variant is real. For example, if results at different time points are available, an increase in the variant frequency over time supports the hypothesis that the variant is subject to selection and thus is real (Hughes et al. 2010).

In the experiments with unmixed viral stocks, excluding observed substitutions in homopolymer runs of four or more nucleotides improved the evidence of purifying selection, as evidenced by decreased π_N:π_S (table 1). This observation is consistent with the expectation that observed substitutions in homopolymer runs include artifacts. On the other hand, the π_N:π_S ratio increased only slightly when rare variants ($<$ 1%) were included but mutations in homopolymer runs were excluded (table 1). The latter observation suggests that many rare variants outside of homopolymer runs are real. Moreover, the estimates of π_S and π_N were substantially higher when rare variants were included (table 1). Thus, if the goal of a study is to estimate population parameters such as synonymous and nonsynonymous nucleotide diversity, the use of a 1% cutoff may be overly conservative and may lead to substantial underestimation of the within-host diversity of viral populations. In studies of genome-wide patterns of nucleotide diversity, the best strategy for obtaining accurate estimates may be to include all observed variants but to mask out homopolymer runs of four or more nucleotides.

Supplementary Material

Supplementary table S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The research was conducted in part at a facility constructed with support from Research Facilities Improvement Program grant numbers RR15459-01 and RR020141-01. S.L.O. and C.M.B. were funded, in part, by a CHAVI/HVTN ESI Award. This research was funded in part by NIH grants GM43940 to A.L.H. and R.F. and AI077376 to David H. O'Connor and by the National Center for Research Resources grant RR000167.

Literature Cited

Ahmadian A, Ehn M, Hober S. 2006. Pyrosequencing: history, biochemistry and future. Clin Chim Acta. 363:83–94.

Allen TM, et al. 2000. Tat-specific CTL select for SIV escape variants during resolution of primary viremia. Nature 407:386–390.

Bimber BN, et al. 2009. Ultradeep pyrosequencing detects complex patterns of CD8+T-lymphocyte escape in simian immunodeficiency virus. J Virol. 83:8247–8253.

Bimber BN, et al. 2010. Whole genome characterization of HIV/SIV intrahost diversity by ultradeep pyrosequencing. J Virol. 84:12087–12092.

Blankenberg D, et al. 2010. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 89:19.10.1–19.10.21.

Bushman FD, et al. 2008. Massively parallel pyrosequencing in HIV research. AIDS 22:1411–1415.

Erickson AL, et al. 2001. The outcome of hepatitis C virus infection is predicted by escape mutations in epitopes targeted by cytotoxic T lymphocytes. Immunity 15:883–895.

Goecks J, Nekrutenko A, Taylor J. Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11:R86.

Guglietta S, et al. 2005. Positive selection of cytotoxic T lymphocyte escape variants during acute hepatitis C virus infection. Eur J Immunol. 35:2627–2637.

Harris RS. 2007. Improved pairwise alignment of genomic DNA [unpublished PhD dissertation]. [University Park (PA)]: Pennsylvania State University.

Hedskog C, et al. 2010. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One 5(7):e11345.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.

Hughes AL, et al. 2010. Dynamics of haplotype frequency change in a CD8+TL epitope of simian immunodeficiency virus. Infect Genet Evol. 10:555–560.

Kajiyama T, Kuwahara M, Goto M, Kambara H. 2011. Optimization of pyrosequencing reads by superior successive incorporation efficiency of improved 2—deoxyadenosine-5′-triphosphate analogs. Anal Biochem. 416:8–17.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing—concepts and limitations. Bioessays 32:524–536.

Langaee T, Ronaghi M. 2005. Genetic variation analyses by pyrosequencing. Mutation Res. 573:96–102.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182:295–301.

Lysholm F, Andersson B, Persson B. 2011. An efficient simulator of 454 data using configurable statistical models. BMC Res Notes. 4:449.

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Martinez-Picardo J, et al. 2000. Antiretroviral resistance during successful therapy of HIV type 1 infection. Proc Natl Acad Sci U S A. 97: 10948–10953.

Moore MJ, et al. 2006. Rapid and accurate pyrosequencing of angiosperm plastic genomes. BMC Plant Biol. 6:17.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nelson EK, et al. 2011. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. BMC Bioinformatics 12:71.

O'Connor DH, et al. 2004. A dominant role for CD8+-T-lymphocyte selection in simian immunodeficiency virus sequence variation. J Virol. 78:14012–14022.

O'Connor SL, et al. 2012. Conditional CD8+ T cell escape during acute simian immunodeficiency virus infection. J Virol. 86:605–609.

Poon AF, 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. 27:819–832.

Quince C, Lanzen A, Davenport RJ, Tumbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12:38.

Rozera G, et al. 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphocyte sub-populations. Retrovirology 6:15.

Truong HM, et al. 2006. Routine surveillance for the detection of acute and recent HIV infections and transmission of antiretroviral resistance. AIDS 20:2193–2197.

Tsibris AM, et al. 2009. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS One 4(5):e5683.

Valentine LE, et al. 2009. Infection with "escaped" virus variants impairs control of simian immunodeficiency virus SIVmac239 replication in *Mamu-B-*08*-positive macaques. J Virol. 83:11514–11527.

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: 1195–1201.

Wang GP, Sherrill-Mix SA, Chang K-M, Quince C, Bushman FD. 2010. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. J Virol. 84:6218–6228.

Wiseman RW, et al. 2009. MHC genotyping with massively parallel pyrosequencing. Nat Med. 15:1322–1326.

Associate editor: Judith Mank