Analysis of unusual and signature APOBEC-mutations in HIV-1 pol next-generation sequences

00528

Philip L. Tzou[1], Sergei L. Kosakovsky Pond[2], Santiago Avila-Rios[3], Susan P. Holmes[4], Rami Kantor[5], **Robert W. Shafer**[1]

[1]Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA; [2]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA; [3]Centre for Research in Infectious Diseases, National Institute of Respiratory Diseases, Tlalpan 4502, 14080, Mexico City, Mexico; [4]Department of Statistics, Stanford University, Stanford, CA, USA; [5]Division of Infectious Diseases, Department of Medicine, Brown University, Providence RI, USA

BACKGROUND

At low mutation-detection thresholds, NGS is susceptible to artifactual detection of mutations from PCR error and APOBEC-mediated G-to-A hypermutation. Here, we analyzed HIV-1 *pol* NGS from published studies to characterize the distribution of unusual mutations and mutations suggestive of APOBEC-mediated G-to-A hypermutation at different detection thresholds.

METHODS

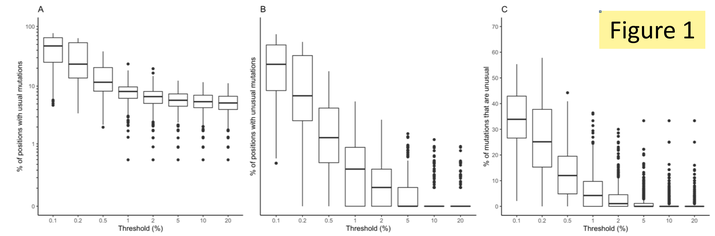
We analyzed published HIV-1 *pol* Illumina NGS data in the NCBI Sequence Read Archive at 8 NGS detection thresholds: 20%, 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%. At each threshold, we determined the % of amino acid mutations that were usual, unusual, or signature APOBEC mutations. Unusual mutations were defined as those with a prevalence <0.01% in the Stanford HIVDB. Usual mutations were defined as differences from subtype B consensus that were not unusual. Signature APOBEC mutations were defined as mutations in APOBEC3G (GG->AG) or 3F (GA->AA) targets at highly conserved positions over-represented in sequences with stop codons and/or active site mutations. Based on a previous study and comparison with LANL Hypermut, *pol* genes with >=3 signature APOBEC mutations were likely to have APOBEC-mediated G-to-A hypermutation.

RESULTS

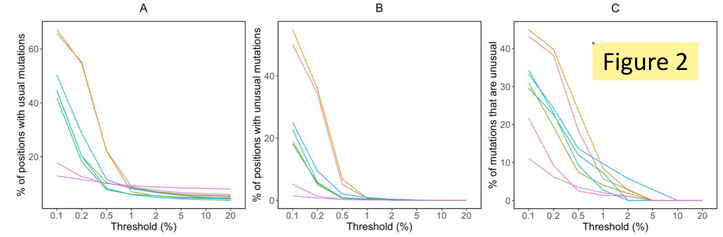
NGS datasets: Eight studies containing 855 samples from 821 persons met study inclusion criteria. These samples included 693 PR, 700 RT, and 449 IN NGS sequence sets. Subtype B accounted for 606 (70.9%) of samples. Subtypes A, C, CRF01_AE, and CRF02_AG were the most common non-B subtypes, accounting for 224 (26.2%) sequences. Four studies included ARV-experienced persons. Plasma HIV-1 RNA levels were available for all samples in 3 studies. Most studies used 0.4 to 1.0 ml of plasma, high-fidelity RT and PCR enzymes, and nested PCR, however, the specific extraction protocols and enzymes used for PCR amplification varied. Across all studies, the median coverage per position was 18,275, with a 5% to 95% range of 2,944 to 81,184.

Pooled data from all datasets: Figure 1 depicts the % of positions with usual (panel A) and unusual (panel B) mutations and the % of mutations that were unusual (number unusual / total number, panel C) as a function of detection threshold in pooled samples for all studies. The median % of positions with a usual mutation increased from 5% to 12% between the 20% and 0.5% thresholds then began doubling to 24% and 48% at 0.2% and 0.1% thresholds. The median % with an unusual mutation increased from 0% to 0.3% between the 20% and 1% thresholds then began increasing by 4-fold to 1.3%, 6.9%, and 23.2% at the at 0.5%, 0.2%, and 0.1% thresholds. The median % of mutations that were unusual increased from 0% to 1.1% between the 20% and 2% thresholds then jumped to 4.2%, 12%, 25%, and 34% at the 1%, 0.5%, 0.2%, and 0.1% thresholds.

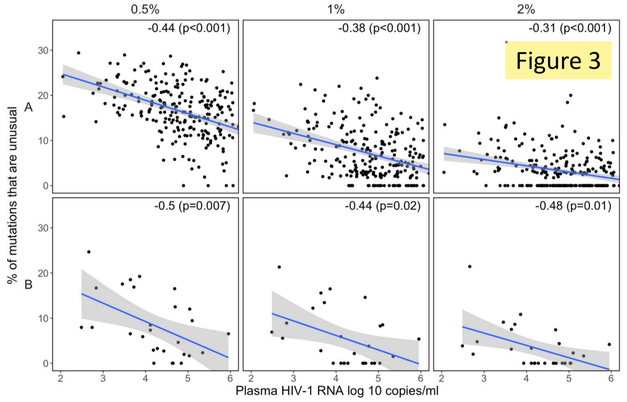
There was a weak significant correlation between the log10 number sequence reads and the number of usual (but not unusual) mutations (r: 0.21-0.24, p<0.001) at the 1%, 2%, 5%, 10%, and 20% thresholds.



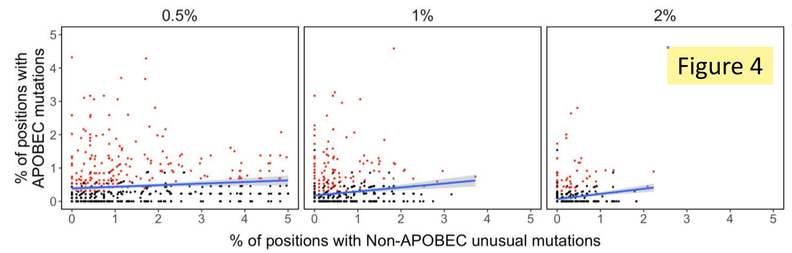
Inter- and intra-study variation: Figures 2A and 2B plot the median % of positions with usual and unusual mutations and 2C plots the median % of mutations that were unusual at the 8 detection thresholds for each study. There were clear inter-study differences, particularly at thresholds <1% with the 8 studies clustering into 3 groups at these thresholds. Because of the inter-study differences in sample characteristics and laboratory procedures, we were unable to attribute the clustering to specific factors. There was also heterogeneity in the distribution of unusual mutations at different thresholds within each study. For example, at the 1% threshold, the highest number of unusual mutations in each study was ~5 times higher than the median number of unusual mutations.



Association of unusual mutations with low virus load (VL): The % of positions with unusual mutations was significantly higher in samples with lower VL in 2 of 3 studies, particularly at 0.5%, 1%, and 2% thresholds (Figure 3). There was no significant relationship between VL and % of positions with usual mutations.



Signature APOBEC mutations: At the 20%, 10%, 5%, 2%, and 1% thresholds, the median % of positions with a signature APOBEC mutation was 0. (The median % began to rise at lower thresholds). However, at the 2% and 1% thresholds, 46 and 113 samples, respectively, contained ≥3 signature APOBEC mutations. Although the complete set of signature APOBEC mutations (n=296) is much smaller than the complete set of unusual mutations that are not signature APOBEC mutations (n=14,940), signature APOBEC mutations outnumbered non-APOBEC unusual mutations in 16.0%, 17.3%, and 13.2% of samples at the 0.5%, 1%, and 2% thresholds (Figure 4). Samples with >=3 signature APOBEC mutations are in red.



CONCLUSIONS

The marked increase in the % of positions with unusual mutations at thresholds <1% and in samples with lower virus loads suggests that, at low thresholds, many unusual mutations are artifactual, reflecting PCR error or G-to-A hypermutation.

In two of the studies for which plasma HIV-1 RNA levels were available, the % of positions with unusual mutations was inversely related to virus load. A likely explanation for this observation is that samples with lower virus loads yield fewer cDNA molecules and that as a result a greater amount of sequence variation results from PCR amplification rather than HIV-1 variation.

The observation that read coverage was not correlated with the % of positions with unusual mutations suggests that (when there is sufficient coverage) most unusual mutations do not result from machine error.

Profiling the numbers of unusual and signature APOBEC *pol* mutations at different NGS mutation detection thresholds may be useful to avoid selecting a threshold that is too low and that poses an unacceptable risk of identifying artifactual mutations.

ADDITIONAL KEY INFORMATION

The work described here has just been published: Tzou, P.L., Kosakovsky Pond, S.L., Avila-Rios, S., Holmes, S.P., Kantor, R. and Shafer, R.W., 2020. Analysis of unusual and signature APOBEC-mutations in HIV-1 pol next-generation sequences. *PloS one*, 15(2), p.e0225352.

We have created a program that performs the analyses described in this study: https://hivdb.stanford.edu/hivdb/by-reads/