Amino Acid Prevalence of Mutations Detectable by Direct PCR versus Single Genome Sequencing

Philip L. Tzou, Soo-Yon Rhee, Robert W. Shafer
Stanford University

Rationale

- We sought to determine if the distribution of amino acids detected by SGS differs from the distribution of amino acids detected by direct PCR population-based sequencing.
- Like NGS, SGS can detect low-abundance variants. Unlike NGS, SGS is not affected by PCR errors. However, SGS can be affected by errors when extracted RNA is reverse transcribed to cDNA.
- In contrast to most other investigators studying SGS, we were not interested in the population genetics of HIV or of the linkage of mutations in the HIV-1 genome.

Rationale

- We believe that understanding the “ecosystem” of mutations detected by SGS, particularly those present at low levels has implications for performing quality control of NGS data.
- Specifically, we hypothesize an excess of highly unusual mutations -- defined as occurring in < 0.01% of viruses by direct PCR Sanger sequencing -- detected at low levels by NGS suggests that these unusual mutations reflect PCR error rather than authentic variants.
- In this scenario, we have suggested that a higher threshold should be used for detecting low abundance variants by NGS.
- Our hypothesis depends in part on the concept that mutations that are highly unusual by direct PCR sequencing are also highly unusual even when they occur at low levels.

Methods: SGS Database

- Using a BLAST search and literature review, we attempted to identify all studies for which multiple clones were sequenced for the same sample using Sanger sequencing.
- Studies using molecular cloning were excluded.
- Studies that contained ≥100 pol SGS were added to a database which is available online at hivdb.stanford.edu/project/sgs.

Methods: Analysis

- Here we analyzed those SGS obtained from plasma in persons with active virus replication (i.e., without virological suppression)
- We determined the HIV-1 subtype and the list of mutations defined as differences from the subtype B consensus sequence.
- We identified which mutations were signature APOBEC mutations.
- We determined which mutations were highly unusual defined as having a prevalence of <0.01% in the Stanford HIVDB and not being a signature APOBEC mutation.

Results

- 24 published studies containing 10,481 pol sequences in 908 plasma samples from 291 persons in GenBank were identified.
- 41% of persons had ≥2 samples.
- Median of 8 SGS per sample (range: 1 to 166).

Gene	# Persons	# SGS
PR	257	8,304
RT	275	9,597
IN	187	4,957

Distinct Mutations in Complete SGS Dataset

Gene	≥ 1 SGS	≥ 2 SGS
PR	325	223
RT	848	605
IN	628	390

Depending on the gene, 29% to 38% of distinct mutations occurred in just one SGS. The possibility that these resulted from RT error cannot be excluded.

62% to 71% of distinct mutations occurred in ≥2 SGS suggesting that these were very likely to be authentic variants.

Number and % of Distinct Highly Unusual Mutations Present in ≥1 and ≥2 SGS

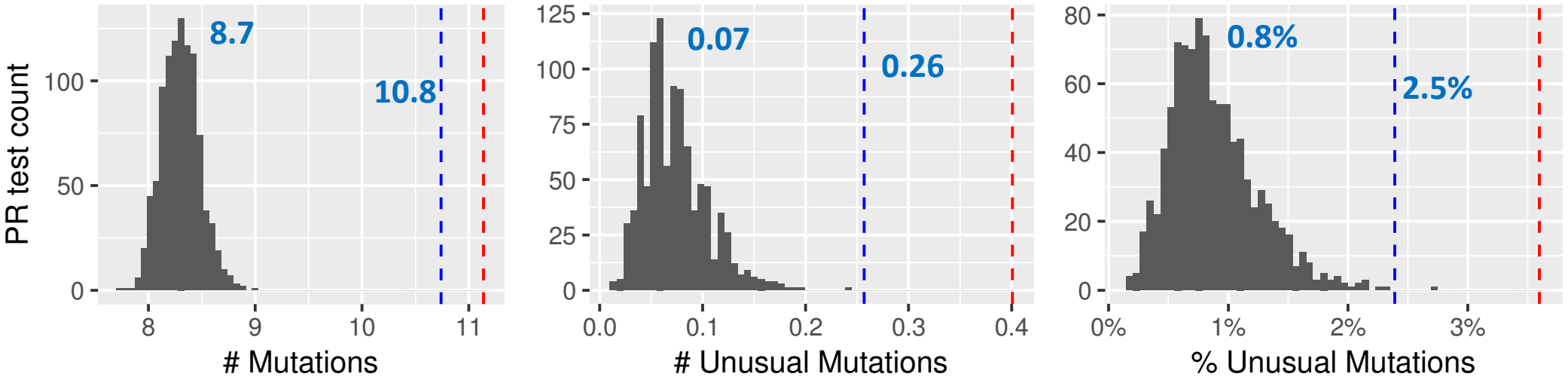
Gene	1 SGS	≥ 1 SGS	≥ 2 SGS
PR	37 (36.3%)	67 (20.6%)	30 (13.5%)
RT	132 (54.3%)	242 (28.6%)	110 (18.2%)
IN	90 (37.8%)	115 (18.3%)	25 (6.4%)

Mutations that occurred in just one SGS were much more likely to be highly unusual compared with those that occurred in ≥2 SGS.

Correlation of mutation prevalence (r^2): SGS vs. population based sequences

	PR	RT	IN
All	0.93_{257}	0.87_{275}	0.84_{187}
Subtype B	0.86_{100}	0.74_{117}	0.65_{52}
Subtype C	0.97_{107}	0.94_{108}	0.91_{108}
Other Subtypes	0.78_{51}	0.80_{51}	0.79_{29}

Distribution of Expected # Mutations, # Unusual Mutations, % Unusual Mutations - PR

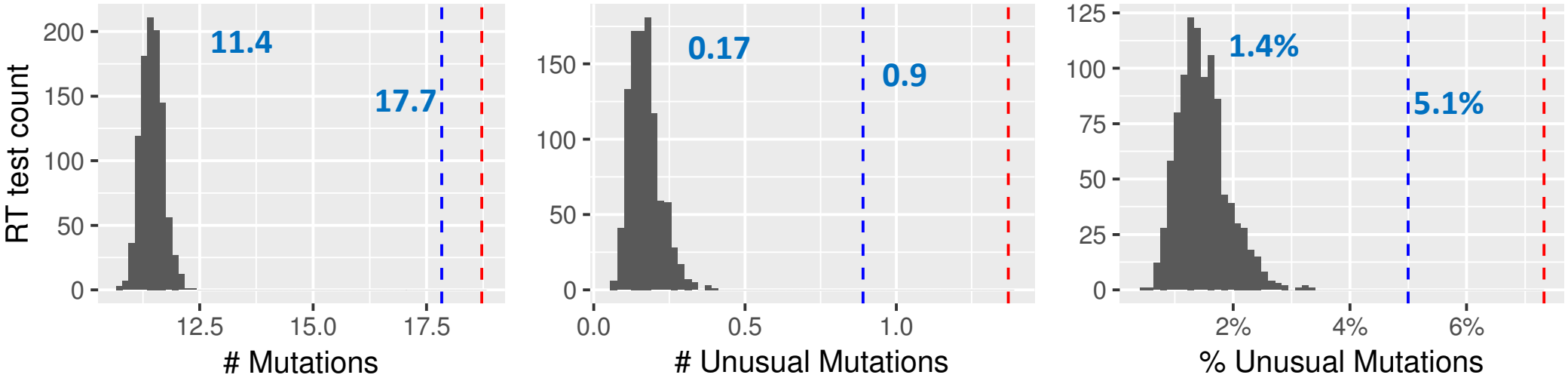


Permutation test: Distribution of # mutations, # unusual mutations, and % unusual mutations in 1,000 samples of PR sequences from 257 persons matched for subtype and ART exposure.

- - - - Mutations in ≥ 2 SGS

- - - - Mutations in ≥ 1 SGS

Distribution of Expected # Mutations, # Unusual Mutations, % Unusual Mutations - RT

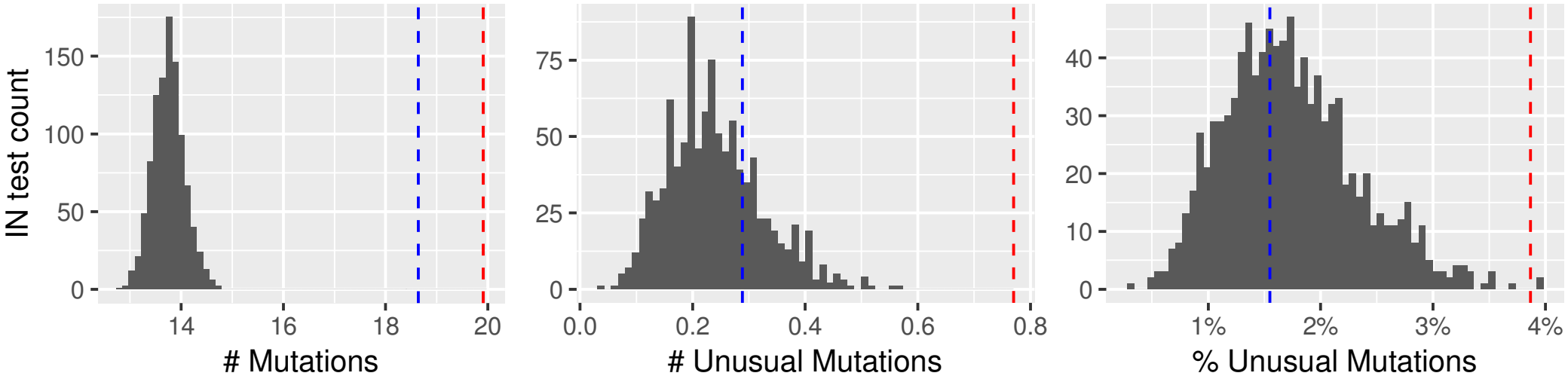


Permutation test: Distribution of # mutations, # unusual mutations, and % unusual mutations in 1,000 samples of RT sequences from 275 persons matched for subtype and ART exposure.

- - - - Mutations in ≥ 2 SGS

- - - - Mutations in ≥ 1 SGS

Distribution of Expected # Mutations, # Unusual Mutations, % Unusual Mutations - IN



Permutation test: Distribution of # mutations, # unusual mutations, and % unusual mutations in 1,000 samples of IN sequences from 187 persons matched for subtype and ART exposure.

- - - - Mutations in ≥ 2 SGS

- - - - Mutations in ≥ 1 SGS

Conclusions

- Highly unusual mutations in PR and RT occurred in higher numbers and higher percentages in SGS compared with population-based sequences.
- In PR and RT, 2.5% and 5.0% of mutations observed by SGS were highly unusual, which was ~3 times higher than what was observed by population-based sequencing.
- This may reflect the fact that SGS detects low-abundance variants that may have reduced fitness and unlikely to reach the levels at which they would be observed by population-based sequencing.
- However, overall there was a strong correlation between the prevalence of mutations observed by SGS and those observed by population-based sequencing.

Conclusions

- This has implications for NGS in that it suggests that a high prevalence of highly unusual mutations is not expected to occur simply because NGS detects low abundance variants.
- This study has two major limitations:
  - The median # reads per sample was just 8.
  - The definition of highly unusual IN mutations may be suboptimal as only about 15,000 population-based IN sequences were available to derive the list of highly unusual IN mutations.