# VQA HIV Gene Sequencing Proficiency Testing Scoring Criteria and Policies

Proficiency at genotypic HIV-1 drug resistance (HIV GEN) testing is assessed by comparing the sequences produced in each participating laboratory with the consensus sequences that are formed by combining data across participating laboratories. Proficiency scoring is limited to the regions of Protease and Reverse Transcriptase (PR and RT respectively) that are used by the International AIDS Society-USA to select mutations that are or may be relevant in evaluating therapeutic efficacy (PR – amino acids 9-91; RT – amino acids 40-237) or the integrase gene (INT – amino acids 50-200). For RT and PR genes, a separate consensus sequence is defined for each sample and commercially available kit (ViroSeq HIV-1 Genotyping System and TRUGENE HIV-1 Genotyping Kit) using all available data sets from a round of testing. The consensus is obtained from a software package that does sequence alignment. A consensus among sequences from internally developed (IH) assays made little sense because of differences among the assays. Therefore, these sequences were scored against a reference that consisted of all of the nucleotide positions at which the consensus sequences from the ViroSeq and TRUGENE kits agreed with each other for that sample and gene. The consensus for INT scoring must be derived from data generated using IH assays because there is no FDA-approved commercial kit for this gene. The consensus sequence was chosen as a basis for proficiency scoring because the proficiency panels consist of clinical samples with unknown true sequences. Proficiency scores are based on the number of disagreements with the consensus sequences (herein referred to as errors). The protease and reverse transcriptase sequences from each sample are scored separately. Scores are summed over the 2 genes and 5 samples for RT and PR and over the one gene and five samples for INT to produce the total score for a laboratory.

The statistical framework for proficiency assessment for gene sequencing is very similar to the approaches that were used to develop other proficiency testing programs under the VQA Program. First, errors are assumed to occur randomly according to a specified probability distribution with parameters that are specific to each sample, gene and kit. Then, the probability of achieving or exceeding an observed error count is calculated from the assumed distribution. Error counts that are associated with low probabilities are flagged as being unlikely to have occurred by chance; i.e. error counts that have low probabilities of occurring by chance will be interpreted as evidence of performance problems.

This general framework is used to score each RT and PR or INT sequence in three stages. First, the total number of errors in a sequence is scored. Second, the total number of Complete Mismatch errors in a subset of types of errors that is defined below is scored. Third, the total number of errors in identifying amino acids at resistance-associated codons is scored. Scoring at the third stage is limited to the resistance-associated codons that are common to the most current mutation libraries associated with both the ViroSeq and TRUGENE assays within the RT and PR genes that are used on a given proficiency panel. For INT, resistance-associated codons were defined by the IAS-USA, March 2013 edition and will be updated as needed.

The subset of errors for performance assessment at the second stage was defined by inspecting the results from early proficiency panels to determine the types of errors made on the panels. The list of observed errors was expanded to include some others that were plausible but had not occurred yet. The following three classes of errors were identified.

1. Partial mismatches. One common type of error is a disagreement between a single nucleotide at a given position in the consensus sequence versus a mixture that included this nucleotide in the submitted sequence (e.g. A vs. IUB code R). The reverse – a mixture in the consensus sequence versus a single nucleotide that is part of the mixture in the sequence from a laboratory – can also occur. More complicated situations, such as a two-part mixture in the consensus versus a three-part mixture that includes the consensus mixture can also occur (e.g. IUB code R vs. IUB code D). All of these are considered partial mismatches.

2. Complete mismatches. The most common example of a complete mismatch is a disagreement between a nucleotide in the consensus sequence and the corresponding nucleotide in the sequence from a laboratory (e.g. A versus G). More complicated mismatches do occur, such as a two-part mixture in the consensus sequence versus a single nucleotide that is not part of the mixture (e.g. R vs. T). An insertion that is reported by a laboratory but not included in the consensus sequence, or a missed insertion that is included in the consensus sequence but not reported by a laboratory, would also be categorized as a complete mismatch.

3. Missing data. In some cases, an entire sequence was missing. In others, part of the sequence was missing. A call of 'N' at a given position in the sequence from a laboratory would also be considered to be missing data if the consensus was something else.

The second stage of scoring is limited to the complete mismatches. The idea at this stage is to evaluate the quality of sequencing given that sequence data were obtained. Partial mismatches are excluded because they may reflect performance characteristics of the kits or minor variants in the sample so additional weight for this mismatch is not warranted. Missing data are excluded so that this stage of the analysis focuses on sequences that were actually obtained.

When proficiency scoring for HIV genotyping was discussed with ACTG virologists, several expressed interest in focusing on resistance codons. This interest motivated the scoring approach described here. The total proficiency score for each sequence in a laboratory is based on a weighted sum of the scores at the three stages. In effect, this approach gives greater weight to complete mismatches than to the other types of errors and greater weight to errors that alter amino acids at resistance-associated codons than to errors that don't alter these amino acids or that occur elsewhere in the sequence.

## Specifying the Consensus Sequence

For purposes of scoring, a nucleotide position is included in the consensus sequence only if there is at least 80% agreement for that position among the laboratories in which a given kit was used. If agreement is less than 80%, then that position is excluded from the performance assessment. Agreement is strictly defined. A call of a single nucleotide in one laboratory and a mixture that includes that nucleotide in another is considered a disagreement. A mixture of two or more nucleotides is considered the consensus only if at least 80% of sequences include the same mixture at that position. Cases that could be considered mixtures, such as those in which, say, half the sequences include one nucleotide at a given position and the other half include a different nucleotide at that position are not considered mixtures for purposes of defining the consensus.

Experience with the first few proficiency panels indicated that very few positions would be excluded from scoring under this approach. Agreement reached at least 97% on both RT and PR genes in each of the five samples over each of the first five proficiency panels. This was true both for nucleotides over the entire sequence expected from each kit and for amino acids at resistance-associated codons. The same approach was adopted for scoring of sequences in the INT gene.

## A Model for Assessing Performance

Under the assumptions that sequencing errors in a given gene on a given sample are both rare and randomly distributed among nucleotide positions and laboratories, the number of errors in the sequence from a laboratory follows a Poisson distribution with unknown parameter $\theta$, where $\theta$ is the expected number of errors per sequence. As noted earlier, a p-value is assigned to each observed error count that represents the probability of obtaining at least the observed number of errors by chance in a random sample from a Poisson distribution with parameter $\theta$. Error counts with p-values below 0.05 or 0.01 are flagged as mild or serious problems respectively.

An estimate of $\theta$ is obtained from the data. Under the assumption of randomly distributed errors, the average error count, across all laboratories in which the same kit was used, is an unbiased estimate of $\theta$. However, performance problems that would result in high error totals could inflate the average error

count, which would bias the estimate of $\theta$. Therefore, the frequency distribution of error counts for each gene in each sample will be examined to determine if there are error counts that are high enough to bias the estimate of the binomial parameter. Under the assumptions made here, these high error counts will appear to be outliers relative to the values that would be expected in a random sample from a Poisson distribution. If such outliers are identified, then the Poisson parameter is estimated after the outliers have been excluded. Otherwise, the parameter is estimated from the complete data.

Inspection of the data from early proficiency panels indicated that bias caused by outliers was an uncommon problem that has relatively little impact on the results of proficiency testing. This conclusion was based on a comparison of the range of error counts that would be flagged in data sets that included one outlier each and the range that would be flagged if the outlier were excluded. For example, suppose the average rate over 15 laboratories was 2.33 errors/sequence and that one sequence from one laboratory included 14 of those errors. Then, the average error rate, excluding this laboratory would be 1.5 errors/sequence. A data set with 5 errors would not be flagged if the parameter estimate included the outlier but would be flagged if the outlier was excluded from the parameter estimate. However, data sets with 4 or few errors would not be flagged in either case while data sets with 6 or more would be flagged in both cases. This is a rather small change in the cut point for determining proficiency. If the outlying data set included 21 errors, then the estimated Poisson parameter after excluding the outlier would be 1.0 and any data set with at least 4 errors would be flagged. Again, this is a small change in the cut point for proficiency testing.

In using the approach to performance assessment that is described here, care must be taken to avoid flagging error counts that are so low that they would otherwise be considered acceptable. Suppose, for example that a set of PR sequences from the ViroSeq kit included one disagreement with the consensus in every three sequences, giving an average error count of 0.33 per sequence. Then, under the Poisson sampling model, a sequence with only two errors in 297 positions (99.3% agreement) would be flagged as having too many errors. Results from Proficiency Panels 002g, 003g, 004g and 005g indicated that this problem was likely to occur. The $25^{th}$ percentile of 0.30 errors per sequence for PR on the ViroSeq kit was very close to the average error count used in the example above. If this value were treated as an estimate of the Poisson parameter then any error count greater than approximately one per sequence would be flagged.

This problem can be avoided if a threshold error rate can be defined such that error rates which are no greater than the threshold will be considered acceptable regardless of the p-value from the Poisson model. A threshold of greater than 1% agreement with the consensus is used. That is, a sequence will be flagged only if the error rate is greater than 1% and the p-value from the binomial model is $\leq$0.05. The threshold error rate of 1% was derived from inspection of results from the coded replicates that were included on panels 002g-005g. In taking this approach, it was assumed that the rate of disagreement with a consensus sequence will generally be equal to or greater than the rate of disagreement between replicates of the same sample within a laboratory; i.e. inter-laboratory variation will generally be at least as great as intra-laboratory variation. Complete intra-laboratory agreement between replicate sequences was relatively uncommon. However, rates of agreement of at least 99% were achieved in the majority of cases for both genes sequenced on both kits. This same rule is applied regardless of the gene being analyzed.

In summary, performance on HIV gene sequencing panels is assessed by assigning a p-value to the observed number of disagreements between the sequence produced in a laboratory and the consensus for that gene in that sample across all laboratories in which the same kit was used. The p-values are derived from a Poisson sampling model under the assumption that disagreements with the consensus are randomly distributed among positions and laboratories. A separate Poisson parameter and separate cut points for assigning p-values is determined for each gene, sample and kit on the assumption that rates of disagreement are likely to depend on all three variables. Error rates $\leq$1% are not be flagged as problematic even if the p-value for such a rate on a given sample is low enough to signal a problem.

Determining the Performance Score

For RT and PR, the results from the three stages of scoring described above are combined to produce a total score by first assigning points for the results to each sequence at each stage and then summing the points across stages, sequences and samples within a laboratory. Points are assigned using the follow algorithm.

| | |
|---|---|
| Stage 1 (Total errors on nucleotide calls): | 2 points if $p \leq 0.01$<br>1 point if $0.01 < p \leq 0.05$<br>0 points if $p > 0.05$ |
| Stage 2 (Complete mismatches): | 1 point if $p \leq 0.01$<br>0 points if $p > 0.01$ |
| Stage 3 (Amino acid calls): | 1 point if $\geq 2$ disagreements with consensus<br>0 points if $\leq 1$ disagreement with consensus |

The resulting scores have a theoretical range of 0-40 (0-4 points per sequence X 2 sequences/sample X 5 samples/panel). For example, complete failure on one of 5 samples would produce a score of 8 points. The total scores are divided into three groups using two cut points. Total scores $\leq 7$ receive performance scores of C, total scores from 8 to 14 receive performance scores of PC and total scores of at least 15 receive performance scores of P.

For INT, the results from the three stages of scoring described above are combined to produce a total score by first assigning points for the results to each sequence at each stage and then summing the points across stages, sequences and samples within a laboratory. Points are assigned using the follow algorithm.

| | |
|---|---|
| Stage 1 (Total errors on nucleotide calls): | 2 points if $p \leq 0.01$<br>1 point if $0.01 < p \leq 0.05$<br>0 points if $p > 0.05$ |
| Stage 2 (Complete mismatches): | 1 point if $p \leq 0.01$<br>0 points if $p > 0.01$ |
| Stage 3 (Amino acid calls): | 1 point if $\geq 2$ disagreements with consensus<br>0 points if $\leq 1$ disagreement with consensus |

The resulting scores have a theoretical range of 0-20 (0-4 points per sequence X 1 sequences/sample X 5 samples/panel). For example, complete failure on one of 5 samples would produce a score of 4 points. The total scores are divided into three groups using two cut points. Total scores < 4 points receive performance scores of C, total scores of ≥ 4, ≤ 7 points receive performance scores of PC and total scores of at least > 7 receive performance scores of P.

After the analysis is complete, a report is sent to each laboratory via email. This summary includes the decoded data from that laboratory as received by the VQA Statistical Analysis Group (SAG), performance on each scoring criterion and a recommended score. The VQA Advisory Board (VQAAB) reviews the analysis and the recommended scores for all laboratories. Please contact Dan Zaccaro regarding any questions with this report (919.541.6310, dzaccaro@rti.org).