Duke Virology Quality Assurance

HIV-1 Genotypic Drug Resistance Sequencing Participation Requirements and Scoring Procedures

Thomas Denny MSc, M.Phil., Principal Investigator
Raul Louzao, MPA Laboratory Operations Director
Miranda Carper, PhD VQA Program Manager
Darin Weed, VQA Program Manager
Salvatore Scianna, QCM Director

VQA Duke Human Vaccine Institute

Contents

VQA Program HIV-1 Drug Resistance Proficiency Testing Requirements

Introduction to Participation Requirements

The National Institute of Allergy and Infectious Diseases (NIAID), Division of AIDS/DAIDS (DAIDS) Virology Quality Assurance (VQA) Program provides quality assurance and proficiency testing (PT) to labs that are performing virologic-based assays for HIV and other pathogens in support of NIAID-funded clinical trials. All laboratories that intend to do HIV-1 drug resistance testing (DR) to identify mutations in the reverse transcriptase, protease and integrase gene regions for NIAID-funded clinical trials must have an Acceptable performance rating in the VQA Genotypic (GEN) HIV-1 Drug Resistance Sequencing proficiency testing program.

Participation in the real-time testing phase of the program can only occur after the successful testing of two five-member panels or completion of an assay validation. VQA GEN HIV-1 DR proficiency testing panels (5 samples) are to be run every six months. The following sections detail the requirements to obtain and maintain an Acceptable performance rating to participate in DR proficiency testing for NIH-funded protocols.

Proficiency Testing Overview:

Beginning January 2023 new scoring rules will be in effect (please see **Scoring Section** below for details). Under the newly approved scoring system sites will now receive a score of either Satisfactory or Unsatisfactory for datasets submitted for each PT event. If a laboratory receives an Unsatisfactory score for a PT, they will receive an Investigation Report (IR) that is to be reviewed and filled out by the laboratory. The IR form helps guide the lab in determining the root cause of the score for that PT. Additionally, there will be two performance rating categories, Acceptable or Not Acceptable. A laboratory must have a Satisfactory grade for 2 out of the 3 past PTs to have an Acceptable performance rating.

The performance ratings are listed in **Table 1**(see the **VQA Program GEN HIV-1 DR Proficiency Testing Scoring Procedure** section below for detailed information on scoring):

Table 1: VQA Performance Ratings

Assessment	PR	PT Score
Acceptable	A	SSS, SUS, USS, SSU
Not Acceptable	NA	UUU, SUU, USU, UUS

PR: Performance Rating
A: Acceptable; NA: Not Acceptable
S: Satisfactory; U: Unsatisfactory

Laboratories enrolled in the Duke VQA Program are shipped 5-member plasma panels twice per year. Laboratories with a WHO-only affiliation receive 5-member plasma panels once per year. Each panel will be named using the following format:

- DNAYYYY_MM.01-05A-E,
  - GEN = PT program
  - YYYY = year of the PT
  - MM = month of PT shipment

- 01-05 = sample number
- A-E = panel configuration

The VQA will provide laboratories with a 5-member PT panel composed of Basematrix 53 Diluent (a defibrinated EDTA plasma produced by SeraCare) spiked with well-characterized HIV-1 virus culture supernatants or virus from infectious molecular clones (IMCs). The viruses will have known HIV-1 drug resistance mutations in the protease (PR; amino acids 9-94), reverse transcriptase (RT; amino acids 40-237) or integrase (INT; amino acids 50-264) genes. Each laboratory must submit a specimen information and mutation file (an EXCEL spread sheet developed by the VQA) and the FASTA/sequence text file for each specimen to the VQA for analyses. Data for the PT is submitted to the VQA for analysis using the Duke VQA Web Application https://vqaapp.dhvi.duke.edu/users/login. Please email the Duke VQA at vqa-gen@duke.edu to register for the web system. The Duke VQA Biostatisticians will analyze data for each PT. Panels will be assessed for sequence homology and mutation calls. Scoring criteria for the VQA GEN HIV-1 DR proficiency testing program is found below in the **Scoring Procedure** section. Scores for each PT will be provided after each round of testing in a report that will be distributed to the participating laboratories via the Duke VQA Web Application.

VQA Participation Requirements

New Laboratory Qualification Testing

Participation in any VQA proficiency testing program must be approved by the VQA Contracting Officer Representative (COR, aka Project Officer). In order to be considered for participation in the VQA Program, a new laboratory must be doing HIV-1 drug resistance testing for an NIH-funded program. Once approved, the new laboratory will need to complete an application for participation in the program that will outline the programs of interest and provide laboratory contact information. A VQA laboratory is identified by a Harmonic ID (HID), Laboratory Data Management System (LDMS), or VQA assigned number.

The VQA GEN HIV-1 DR proficiency testing program currently evaluates assay performance for RT, PR, and INT gene regions. RT and PR genes are evaluated together, and the INT gene region is evaluated separately. Laboratories may opt to participate in one or both programs. Participation rules are the same regardless of the gene region being evaluated.

The VQA program accepts data for genotypic HIV-1 drug resistance testing generated using commercially available and internally developed assays (IH). Sequence files must be submitted to the VQA via the Duke VQA Web Application https://vqaapp.dhvi.duke.edu/users/login. All new laboratories must register with the Duke VQA. To register, email the Duke VQA at vqa-gen@duke.edu in order to gain access to the VQA Web Application. Information regarding panel labeling requirements and rules for data submission will be provided to the testing laboratory through the VQA Web Application and/or email when panels are shipped for testing.

A new laboratory must obtain an **Acceptable** performance rating prior to performing any protocol testing. In order to achieve an Acceptable performance rating for genotypic HIV-1 drug resistance testing, a laboratory must run an assay validation (VQA-provided or in-house) and pass two (2) consecutive or 2 of the most recent 3 five-member panels. The testing may be done during normal track testing (panels will be received every six months), or fast-tracked using historical panels with previously derived consensus sequences. In order to identify problems early in the process, first 5-member panel must be completed and scored prior to testing the next set of samples.

The final determination of whether or not a laboratory may perform protocol testing is at the discretion of the program leadership for that laboratory, not the VQA.

New Assay Enrollment and Qualification

A laboratory may request to:

- Qualify additional assays for protocol testing
- Switch platforms
- Validate additional instruments for qualified assays

All new assays and instruments enrolled in the VQA program need to be qualified following the procedures outlined in the above section, **New Laboratory Qualification Testing.** The VQA permits the use of two different assays/platforms; laboratories must obtain special approval from the VQA if they wish to become qualified using three or more different assays. Each assay added to the VQA Program will be qualified separately and the rules for achieving an Acceptable performance rating applies to all assays used within a laboratory.

The Duke VQA can assist in assay and instrument validations. Please contact the Duke VQA (vqa-gen@duke.edu) for more information about assay requirements or to request a validation plan and data submission template.

Maintaining an Acceptable Performance Rating

To maintain an Acceptable performance rating, a laboratory must do the following:

- Participate in two rounds of real-time proficiency testing per year as scheduled by the VQA.
- A laboratory must have a Satisfactory score for 2 out of the 3 recent rounds of proficiency testing.

Extensions will no longer be granted to sites. PT due dates will be adjusted as needed if the receipt of a shipped PT panel at a participating laboratory is delayed. If a lab does not submit data by the PT due date the lab will receive a score of Unsatisfactory. However, a laboratory may request a PT exemption if they are unable to participate in proficiency testing due to circumstances out of their control. All exemption requests must be approved by the VQA prior to the PT due date. Acceptable reasons to temporarily be exempt from proficiency testing include:

1. Force Majeure (hurricane, tornado, flood, fire, etc.)
2. Laboratory closures for emergency circumstances (fire, pandemic, radiation leak, flood, electric issues, *etc*.)
3. Supply chain issues with vendor
4. Broken instrument
5. Government shutdown or political unrest

Exemption requests should be made to vqa-gen@duke.edu prior to the start and/or close of the PT event**.** A laboratory may receive a score of Unsatisfactory if the reason for not submitting results does not meet one of the acceptable criteria for an exemption. **A laboratory must contact their network leadership if they are exempted from a round of VQA GEN HIV-1 DR proficiency testing.**

Withdrawal/Removal

A laboratory may voluntarily withdraw from the VQA GEN HIV-1 DR proficiency testing program at any time. A laboratory that has not participated in the VQA GEN HIV-1 DR proficiency testing program for more than 12 consecutive months will automatically be removed from the program.

Requalification

A laboratory will have to requalify as a new laboratory (see section above) if:

- A laboratory wishes to re-enter the HIV-1 drug resistance proficiency testing program after removal or has not participated in longer than 12 months;
- Laboratory has ongoing problems with proficiency panels

"B-E" panel configurations will be created for each round of testing to permit retesting in the event of problems. A laboratory will need a Satisfactory score on two (2) consecutive or 2 of the most recent 3 five-member panels. If a laboratory receives a Not Acceptable performance rating, they may run a "B" panel configuration for the most recent panel that failed. Additionally, other archived panels may be tested if a laboratory wishes to requalify before the next scheduled panel send-out.

Appeals

The VQA recommends scoring for proficiency panels based on the criteria defined for the program. The VQA Advisory Board (VQAAB) then reviews the scoring for each round of testing for labs that had a decline in status or had a problem that does not follow the VQA PT rules (the laboratory identities are blinded for this process). Any laboratory may appeal the score on a proficiency panel by submitting a letter or email to William Meyer III, Chair of the VQAAB (william.a.meyer@questdiagnostics.com). All appeals will be reviewed by the VQAAB to determine if a change in scoring is indicated. Laboratories will be notified of the outcome of all appeals.

Practice Panels

Any laboratory may obtain retrospective HIV drug resistance proficiency panels prior to enrollment in the program, or at any time during their participation in the VQA program. Once a panel has been tested and analyzed as a 'practice' panel, it cannot be reclassified for use in laboratory qualification. Results from these 'practice' panels will be assessed and results returned to the laboratory, but no score will be assigned.

Change in Status Letters

A laboratory will receive a change in status letter if they obtain a score on a round of testing that changes their overall performance rating. This letter will document the laboratory's scores over the last three rounds of testing and will indicate when a change in status (performance rating) has resulted. A copy of this letter will be sent to the director of the laboratory and the network laboratory group for whom the laboratory does testing, as appropriate. Letters will be sent to notify individuals of both negative and positive changes in performance ratings. The VQA submits the letters on behalf of the VQAAB, but has no control over the implementation of rules governing the ability of a site to continue protocol testing. All questions surrounding a laboratory's ability to resume or discontinue protocol testing should be directed to the respective network laboratory group or leadership.

VQA Program HIV-1 Drug Resistance Sequencing Proficiency Testing Scoring Procedure

Introduction to PT Scoring

Proficiency at genotypic HIV-1 drug resistance 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 9-94; RT – amino acids 40-237) or the integrase gene (INT – amino acids 50-264). For the RT, PR and INT gene regions, a separate consensus sequence is defined for each sample using all available data sets from a round of testing. The consensus is obtained from a program developed by the Duke VQA biostatistics team.

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.

Scoring Criteria for VQA HIV-1 Drug Resistance Proficiency Testing

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.

Technical Performance: Three stages of Scoring

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 within the RT, PR, and INT genes that are identified in the most current edition of the IAS-USA guidelines.

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.

Classes of Errors for Stage II of Scoring: 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.

Classes of Errors for Stage II of Scoring: 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.

Classes of Errors for Stage II of Scoring: 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. 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 θ, where θ 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 θ. Error counts with p-values below 0.05 or 0.01 are flagged as mild or serious problems respectively.

An estimate of θ 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 θ. However, performance problems that would result in high error totals could inflate the average error count, which would bias the estimate of θ. 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 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 25th 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 <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 <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. 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 for PR/RT and INT using the algorithm summarized in **Table 2**.

For PR and RT gene regions, 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 (see **Table 3**).

- Total scores ≤7 receive performance scores of Satisfactory;
- Total scores from 8 to 14 receive performance scores of Satisfactory, however the laboratory will also receive a Potential Issue Alert (PIA) or an information notice;
- Total scores of at least 15 receive performance scores of Unsatisfactory

For INT gene region, 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 see **Table 3**.

- Total scores < 4 points receive performance scores of Satisfactory;

- Total scores of 4 to 7 points receive performance scores of Satisfactory, however the laboratory will also receive a PIA.
- Total scores of at least > 7 receive performance scores of Unsatisfactory.

Table 2: Genotypic HIV Drug Resistance Proficiency Criteria and Targets

Criteria	Description	Performance Score
Total Nucleotide Errors	Total errors on nucleotide calls as compared to the consensus sequence	2 points if p≤0.01 1 point if 0.01<p≤0.05 0 points if p>0.05
Mismatches	**Partial Mismatches** Submitted sequence mixture includes a disagreement in a single nucleotide at a given position when compared to the consensus	1 point if p≤0.01 0 points if p>0.01
	Complete Mismatches* Submitted sequence includes a disagreement between a nucleotide in the consensus sequence	
	Missing Data Full or partial sequence is missing from submission	
Amino Acid Calls	Total errors in identifying amino acids at resistance-associated codons	1 point if ≥2 disagreements with consensus 0 points if≤1 disagreement with consensus

*insertions and deletions will receive the maximum penalty of four errors, reflecting the large amount of mismatches caused by the subsequent frameshift

Table 3: Proficiency Scoring

GEN PR/RT Data Score Range	GEN INT Data Score Range	Scoring
0-7	0-3	Satisfactory (S)
8-14	4-7	Satisfactory (S) +PIA
15-40	>7	Unsatisfactory (U)
Phylogenetic Tree Misclustering	Phylogenetic Tree Misclustering	Unsatisfactory (U)

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 Duke VQA, performance on each scoring criterion, and a recommended score. Please contact the Duke VQA (vqa-gen@duke.edu) regarding any questions with this report.

Additional Scoring Criteria

Below is a list of other criteria that impact a laboratories score:

1. Phylogenetic Tree Misclustering

A phylogenetic tree analysis is performed using all Participant PT data to identify contamination, upload of the incorrect sample data, or other issues. If a site's data for a sample does not cluster with the other PT data, a site will receive a score of Unsatisfactory (see **Table 3**).

2. Data Timeliness

A due date is assigned to each 5-member panel. Data must be submitted via the Duke VQA Web Application by the deadline or the lab will receive an Unsatisfactory Score for the PT. Please see **Maintaining an Acceptable Performance Rating** section above for acceptable reason to request a PT exemption.

3. Data Entry Errors

Data for each HIV-1 Drug Resistance PT is submitted using the VQA web-system, which requires upload of FASTA files for each PT sample and a Specimen Mutation File. Participating sites have the ability to save their data in the web-system and review it before submission. Prior to the PT due date, a site may request to re-open their PT to correct data entry errors. However, sites will not be able to correct any data entry errors after the PT due date. Data that is submitted to the VQA that contains data entry errors such as entering the wrong values, entering data in the wrong location, switching samples, or submitting the wrong file will receive a score of Unsatisfactory.