



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Statistical Analysis of The Amerithrax Morph Assays: Fat Tails and the "False Negative" Rate

S. P. Velsko

November 10, 2015

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Statistical Analysis of the Amerithrax Morph Assays: Fat Tails and the “False Negative” Rate

Steve Velsko
Lawrence Livermore National Laboratory
December 2, 2013
LLNL-TR-679142

Quantitative copy number data reported for several of the Amerithrax morph assays are not consistent with Poisson sampling statistics, but instead exhibit “Taylor’s Law” behavior where the variance greatly exceeds the average. A plausible statistical model for this behavior may explain the large number of observed negative and “inconclusive” findings, and implies that the probability that a repository sample could contain a morph at concentrations well above the nominal detection limit but nonetheless give a negative or inconclusive test result far exceeds what would be predicted based on Poisson statistics.

This document was initially released with restricted distribution as LLNL-TR-647249, but has been approved for unlimited distribution by the Government Accountability Office, for whom it was originally prepared.

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Introduction

The National Academies report on the scientific evidence used in the Amerithrax case questioned the probative value of the so-called “morph assays” for establishing that RMR-1029 was the only source for the material sent in the letters.¹ However, their criticism primarily centered on the possibility that not every possible isolate was, in fact, submitted to the FBI repository or that some isolates may not have been prepared correctly for submission. In addition, the committee pointed out the need to explore the possibility that a morph that was present at a concentration below the detection limit of its assay might be amplified to detectable levels under culture conditions similar to those used to grow the material used in the letters. They did not, however, question the performance characteristics or interpretation of the morph assay results *per se*.

Nonetheless, one of the salient questions that the FBI needed to address to establish the probative value of the assay results was how likely it was that a repository isolate could test negative for a morph even though the morph was, in fact, present at a level above the nominal detection limit of the assay. More generally, what was the “false negative” rate of the morph assays? Although this was critical for assessing the ability of the assays to exclude isolates as potential sources of the letter material, neither the FBI nor the contracting laboratories performing the assays attempted to establish this through a deliberately designed experiment.

However, this question was partially addressed as part of the assay characterization process by testing thirty replicate samples drawn from RMR-1029, a batch of material that was known to contain all four signature morphs at concentrations similar to those in the letters.¹⁻³ A surprising result was the large number of “negative” or “inconclusive” assay results for some of the assays. If these empirical results are used as estimates of “false negative” detection rates of the assays, it is straightforward to calculate the probability that the assays “missed” repository samples that contained all four morphs.²

The interpretation of the morph assay results for the RMR-1029 replicates is complicated by the fact that the “negative,” “positive” and “inconclusive” detection findings reported for these PCR assays are the result of conventions adopted by the assay developers. Each developer established a nominal “limit of detection” (LOD) for his/her assay, implying that a negative finding is not an indication that the morph is absent, but only that it is present below the LOD. The LOD value varied among the assays. The convention for declaring an “inconclusive” result varied among assays as well, but was generally associated with “partial” findings in which not all replicate subsamples drawn from a repository sample tested negative or positive.

The purpose of this report is to examine in some detail the consequences of these conventions on the interpretation of the morph assays. In particular we are

interested in explaining the large number of observed negative and “inconclusive” findings among the RMR-1029 replicates, and exploring the basis for estimating the probability that a repository sample could contain a morph at concentrations well above the nominal detection limit but nonetheless give a negative or inconclusive test result. In the next section we briefly describe the assay characterization process and interpretation criteria used by the various assay developers. An important statistical issue is the sampling variance expected from nominally replicate samples. Subsequently we examine quantitative PCR data provided by some of the developers and show that sampling fluctuations in copy number are not in agreement with Poisson statistics as is often assumed for PCR. We describe a plausible model to explain this. In the following section we explore the consequences of this non-Poisson behavior on the “false negative” detection probability. We conclude with some observations about the FBI’s use of statistical analysis in the Amerithrax case.

Morph assay results depend on copy number fluctuations

A chain of several steps separated the *Bacillus anthracis* Ames stocks held by the various laboratories in question and the samples tested with the morph assays. First, isolates were prepared for the Ames repository from sub-samples of a laboratory’s stocks. Subsequently, samples thought to be representative of each isolate from the repository were sent to the laboratories performing the morph assays.⁴ Each assay development laboratory used a similar procedure to prepare aliquots for PCR analysis, shown schematically in Figure 1. The concentration of DNA was adjusted so that each aliquot of Ames DNA added to the PCR reaction contained approximately 10^7 genome copies, a mixture of wild type and variant sequences.

The FBI specified that each assay provide positive results if the relevant variant sequence were present at the 0.1% level relative to the Ames wild type sequence. Since the total number of copies was $\approx 10^7$, this implied a minimum detection limit of $\approx 10^4$ variant genome copies per PCR reaction. In fact, the assays exceeded this specification and gave positive results for variant fractions between 0.01 and 0.001%, corresponding to 100 – 1000 variant genomes per reaction.

A parent sample (and *by implication* the original repository sample) was declared “positive” for a morph if all three PCR replicates gave above-threshold signals with Ct values less than a pre-decided cutoff value, usually around 40 cycles. If all three replicates did not cross the detection threshold for Ct less than the cutoff value, the parent sample was declared “negative”. Laboratories differed on how they treated the case where only one or two of the replicates were PCR positive. In some cases this could apparently lead to an “inconclusive” finding, while in others a decision was made on the basis of testing of additional replicates.

There are two ways that a PCR assay can give a negative result in a single replicate. A negative result might occur because of contamination of that replicate with an inhibitor or because some reagent was mistakenly left out. However, the more likely explanation is that the morph is present at a concentration lower than the nominal detection limit of the assay as determined by the cutoff Ct. In the rest of this report we will be concerned with this latter mechanism. It is useful to define an “apparent negative rate” (ANR) as the fraction of replicates that would test negative if a large number of replicate sub-samples from the parent sample were tested by the morph assay.

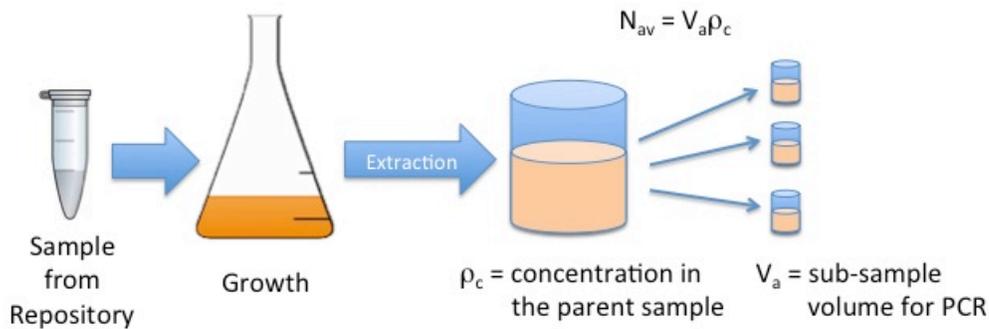


Figure 1. Samples received by the assay performers were cultured, and DNA was subsequently extracted from the cultured material. The resulting DNA suspension is referred to as the “parent” sample. Sub-samples of the parent sample were added to the PCR assay reaction mixture in triplicate. The total number of copies of Ames DNA in the reaction mixture was approximately 1×10^7 .

By its nature, PCR is a “counting” technique because in the absence of significant inhibition the Ct value that obtains when the fluorescence signal crosses the threshold value is a direct reflection of the number of copies of target DNA originally present in the sample. The ANR is thus a reflection of the *variance* in copy number across the replicates drawn from the parent sample. This variance is an important statistical quantity because it ultimately determines the probability that a parent sample can be misclassified as “positive” or “negative” on the basis of the behavior of the replicates. Therefore, in the next section we will assess what the data gathered by the assay developers says about the variance in copy number associated with the morph assays.

Morph assay PCR data implies that copy number fluctuations are not Poisson distributed

The standard model for copy number fluctuations assumes that the DNA molecules suspended in the parent sample are independently distributed in the sample volume, and the probability of capturing k molecules (copies) in any sub-sample is governed by the Poisson distribution:⁵

$$P(k|\rho_v, V_a) = \frac{(\rho_v V_a)^k e^{-\rho_v V_a}}{k!}$$

where ρ_v is the density of variant genomes in the parent sample and V_a is the sub-sample volume. The average number of variant DNA copies contained in V_a is given by $N_{av} = \rho_v \cdot V_a$. An important property of the Poisson distribution is that the expected variance σ^2 is equal to the mean: $\sigma^2 = N_{av}$, and for N_{av} larger than 100, the probability of obtaining k outside the interval $N_{av} \pm 2\sigma$ is very small.

While the results were reported as qualitative categories “positive”, “negative” or “inconclusive”, the morph assays were actually quantitative in nature, and several laboratories reported Ct calibration data and copy number estimates for a variety of samples. A simple test for Poisson behavior is to plot the empirical variance in copy number against the empirical mean copy number μ on a logarithmic scale for replicate samples. For triplicate PCR reactions performed under identical conditions giving copy number estimates n_1 , n_2 , and n_3 :

$$\mu = \frac{1}{3} \sum_{j=1}^3 n_j \quad \text{and} \quad \sigma^2 = \frac{1}{3} \sum_{j=1}^3 (n_j - \mu)^2$$

The results of such a plot are shown in Figure 2.

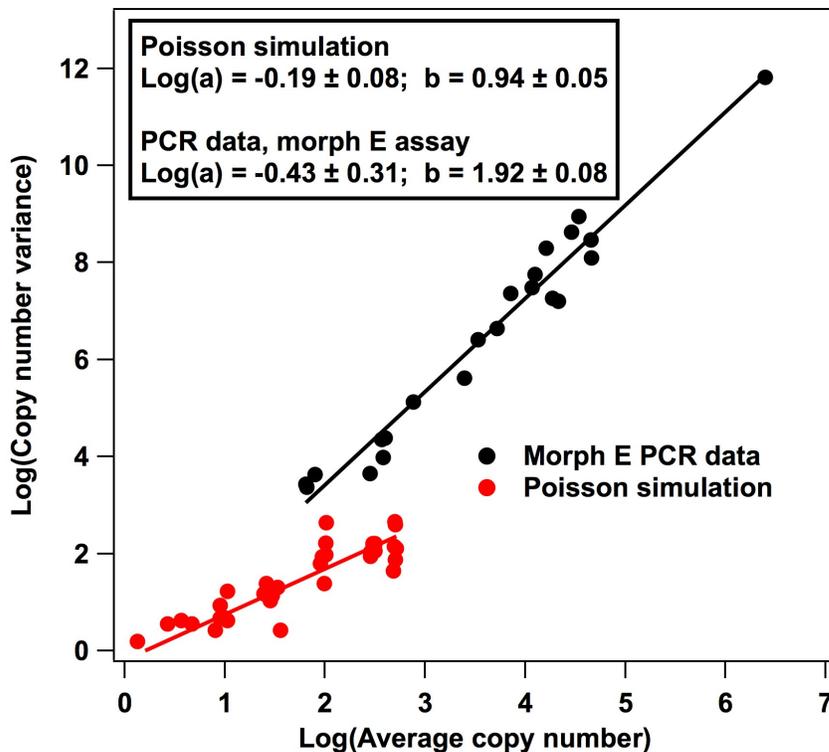


Figure 2. Observed relationship between the average copy number and variance for the TIGR PCR data (Morph E assay) and a simulated Poisson process.

The data in Fig. 2 are derived from copy number data for the E morph assay. For comparison Fig. 2 also shows data from a simulation of Poisson distributed random numbers. The data shows clear evidence of deviation from Poisson behavior $\sigma^2 = \mu$.

The morph E copy number fluctuation data fits reasonably well to a power law relationship between the mean μ and variance: $\sigma^2 = a\mu^b$ with $b > 1$. This relationship is known as Taylor's Law, and was originally observed in ecological data describing the spatial distribution of organisms.⁷ Taylor recognized that the power law phenomenon was generally associated with the fluctuations of the density of units that tend to aggregate rather than move independently. A past study has demonstrated that PCR copy number fluctuations from a wide variety of data sources obey Taylor's law.⁸

IITRI provides similar data for the morph D assay, but their calibration is in nanograms rather than copy number.⁹ However, this only affects the $\text{Log}_{10}(a)$ intercept of the linear least-squares fit and not the exponent b , which is the critical indicator of Taylor's law behavior. Plotting and fitting this data leads to exponents between 1.9 and 2.3 so it appears that the copy number fluctuations for the D assay also obey Taylor's Law rather than Poisson statistics. While there is similar data for the Morph A1 and A3 assays, it exhibits an unfortunate artifact that amounts to a bias – every third replicate has a higher Ct, evidently because some internal standard was added to that PCR reaction mixture for calibration purposes, and it affects the PCR reaction.¹⁰ MRI does not provide the copy number information needed to assess their D morph assay.¹¹

Copy number fluctuation “fat tails” imply larger probabilities of negative and inconclusive findings for “positive” parent samples

One possible mechanism for Taylor's law behavior in PCR reactions is aggregation of target DNA. A model for the sampling statistics of units that aggregate has been developed by Kendal.¹² This distribution cannot be written in closed form, but must be evaluated using numerical techniques. In the following discussion we use the Kendal distribution with a and b parameters equal to that observed for the TIGR data (Fig. 2) to calculate the probability of observing N DNA copies in a sampling volume. The Kendal distribution has “fat tails” in both the low copy number and high copy number regimes compared to the Poisson distribution, and this makes sampling events with extreme values much more probable.

To illustrate the potential implications for the ANR, we have compared the probability of observing N DNA copies in a sub-volume sampled from a parent sample that contains an average of 100 copies per sub-sampling volume V_a . Referring to Fig. 1, this assumes that the density of variant DNA copies in the parent volume is such that $\rho_v V_a = N_{av} = 100$.

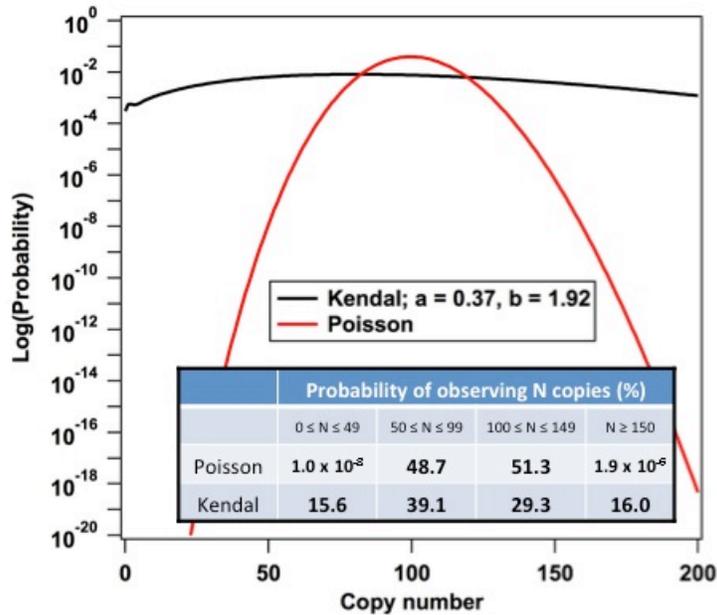


Figure 3. Poisson probability for observing N copies versus the Kendal distribution having the same a and b parameter values as the TIGR data. N_{av} is assumed to be 100 for both distributions.

The inset table in Fig. 3 illustrates the large probability of obtaining low copy ($N < 49$) sub-samples compared to predictions from Poisson statistics. If we assumed that Poisson statistics applied to the TIGR samples we would predict that it is nearly impossible ($P \approx 10^{-8}$) to obtain a sample with $N \leq 49$ if $N_{av} = 100$. However, if the Kendal distribution is a more accurate descriptor of the statistical fluctuations consistent with the observed Taylor behavior, there is actually a $\approx 16\%$ chance of observing such low copy number events. To the extent that these low copy number events are associated with apparent negative detections, its clear that the ANR would greatly exceed what would be expected if Poisson statistics applied.

Also note that if we defined a “positive” sample as one with 150 or more copies per V_a of morph E in 10^7 copies of wild type, then a sample with only 100 nominally would be classified as a “negative” sample. However, such a sample has a 16% chance of contributing 150 or greater copies per sampling volume, clearly a “false positive” event by definition. Thus, if the Kendal distribution describes the copy number fluctuations of the morph assays, the fraction of “high copy number events” is also much higher than would be predicted if Poisson statistics are assumed.

Suppose that we have a case where $C_t=40$ is taken as the cutoff, and is known to correspond to 100 copies per sampling volume in the parent sample. Approximately 50% of the replicates drawn from this parent sample will give $C_t > 40$ and $\approx 50\%$ will give $C_t \leq 40$. Thus, about half the replicates will test “positive” and half will test “negative”. What is the probability that a replicate drawn from a different parent sample containing N_{av} copies per replicate sampling volume V_a will contain *fewer* than 100 copies, and therefore give $C_t > 40$ – a negative result? Conversely, what is the probability that a replicate drawn from a different parent

volume containing N copies per replicate sampling volume will contain *greater* than 100 copies, and therefore give $C_t \leq 40$ – i.e. a positive result? Figure 1 shows these probabilities as a function of N_{av} when the fluctuations in copy number are governed by Poisson and Kendal distributions respectively.

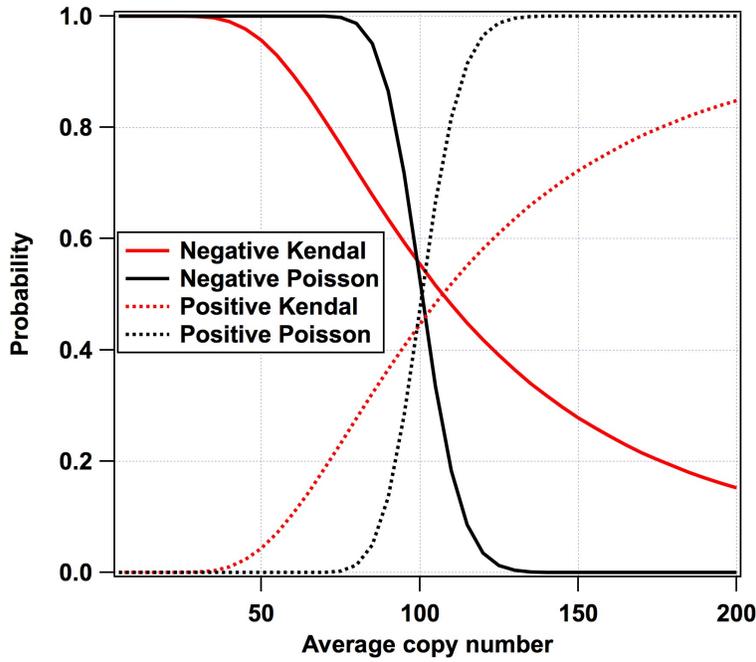


Figure 4. Probability of obtaining negative and positive results for a single replicate subsample drawn from a parent sample containing the indicated average number of copies of DNA per subsample volume. This figure assumes that a C_t cutoff was determined for a parent sample containing an average of 100 copies per subsample volume.

Fig. 4 indicates that if the Kendal distribution governs the fluctuation statistics, a single replicate drawn from a parent sample that contains only 50 copies per subsampling volume on average still has a few percent chance of giving a positive result (red broken line in Fig. 4), whereas if Poisson statistics obtain, the chance of a positive is vanishingly small ($P \approx 10^{-10}$). Similarly, under Poisson statistics, it is essentially impossible to get a negative result if $N > 140$, but Kendal statistics will result in very substantial “false negative” probabilities.

Suppose that the decision to label an assay result as positive or negative is made on the basis of the results for *triplicate* samples. This can be derived from the data in Fig. 4 by using the formula $P(\text{triplicate}) = P(\text{single})^3$. Figure 5 shows the probability of getting all three samples negative or all three samples positive under the same conditions as Fig. 4.

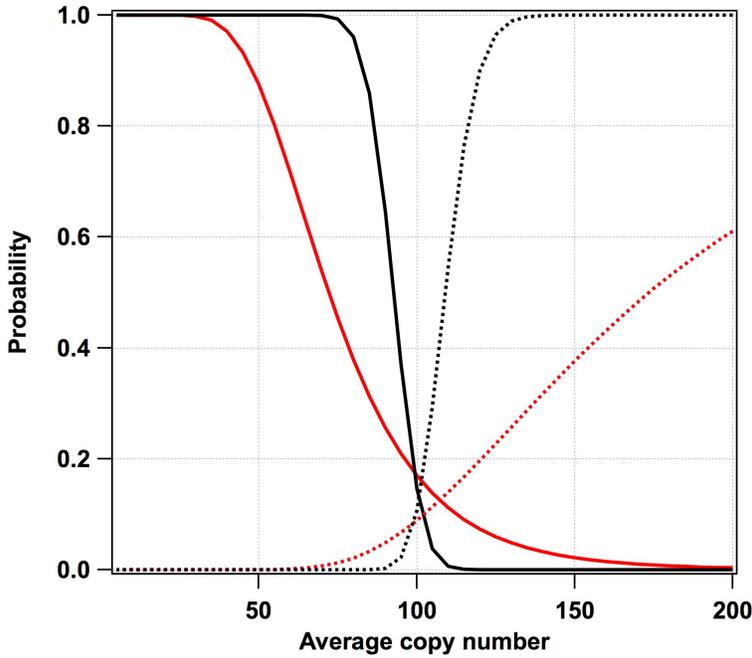


Figure 5. Probability of obtaining negative and positive results for all three replicate subsamples drawn from a parent sample containing the indicated average number of copies of DNA per subsample volume. This figure assumes the same conditions as Fig. 1. Black – Poisson; Red – Kendal; Solid: Negative detection (all three); Broken: Positive detection (all three).

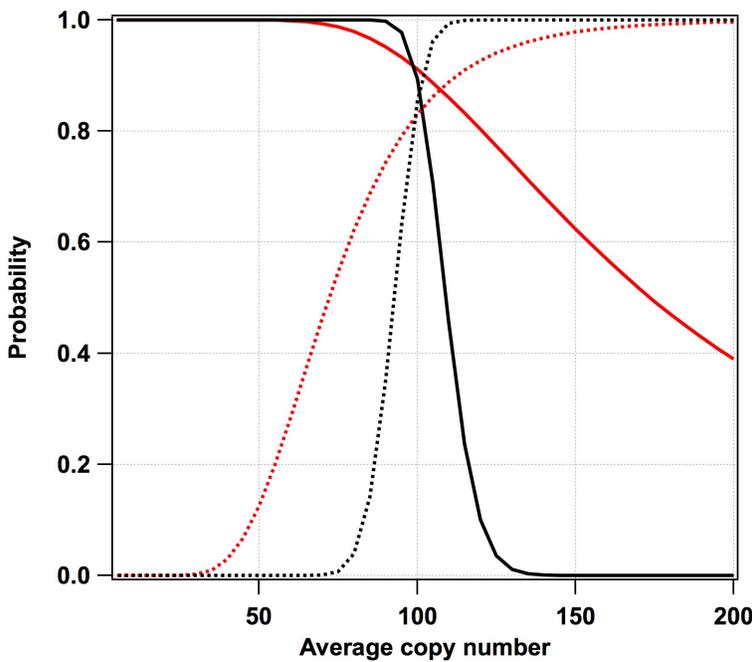


Figure 6. Probability of obtaining negative and positive results for at least one of three replicate subsamples drawn from a parent sample containing the indicated average number of copies of DNA per subsample volume. This figure assumes the same conditions as Fig. 1. Black – Poisson; Red – Kendal; Solid: Negative detection (at least one); Broken: Positive detection (at least one).

The curves in Fig. 5 indicate that under Kendal statistics, more than 10% of triplicates will not be all negative even though the parent sample has only 50 copies per subsampling volume on average. Similarly, only $\approx 40\%$ of triplets will be all positive even though the parent sample has an average of 150 copies per sampling volume. Thus, the Kendal model predicts a much wider region of “inconclusive” results where the three replicates are not either all positive or all negative compared to the case where Poisson statistics apply. This conclusion is reinforced by Fig. 6, which shows the probability that at least one of three replicates will test positive or negative as a function of copy number in the parent sample. For this figure, the probability of at least one positive or negative event is calculated by the formula $P(\text{at least 1}) = 1 - (1 - P(\text{single}))^3$.

Conclusions

“... the FBI did not address the issue of false negative results. In connection with this issue of sensitivity of the assays, a major concern regarding the Statistical Analysis Report is the restriction of its analyses to the 947 samples that contained no inconclusive or variant results.”

National Academies Report¹

The limitations of the FBI’s assay validation effort can be best understood using the concepts of “analytical” and “inferential” validation. Analytical validation demonstrates that the assay can reliably detect the presence of the morph at or above a certain concentration, in a background of wild type. The blind tests administered to the assay development laboratories provided adequate demonstrations of this. However, the assay test results *per se* were only relevant evidence insofar as they supported or refuted the ultimate inference that a lab was a potential source of the letter material. This inference rested, in turn, on the assumption that a lab could be excluded as a source of the *B. anthracis* used to make the letter material if its Ames isolates *tested negative* for one or more morphs found in the letters. The FBI did not show that this inference was empirically valid. Thus, while they were quick to recognize the potential importance of the morph assays, the FBI appeared to adopt a somewhat simplistic picture of how those assay results should be interpreted.¹³ A negative assay result was effectively taken to mean that the morph was absent, and therefore could not generate a spore powder containing that morph.

Instead, a careful analysis of the quantitative PCR data accumulated during the development and validation of the morph assays reveals that the problem of “false negatives” is indeed a concern since samples that contain variant concentrations significantly higher than the nominal LOD can nonetheless exhibit high rates of negative or inconclusive detections.

In this regard it is not clear how the FBI defined its requirement for a minimum LOD of 0.1%. Suppose that samples from several laboratories tested positive for all four

morphs according to the assay criteria. Did they intend to compare the quantitative morph concentrations in order to identify a closer match to the letter materials? In the end, they did not have to confront this problem because one of the two laboratories that had samples in the repository that tested positive for all four morphs could be excluded on the basis of other evidence.

Finally, the calculations presented above illustrate the fundamental problem with setting an arbitrary criterion for a “limit of detection” in PCR assays and reporting results as “positive”, “negative”, or “inconclusive”. If we say that the LOD is a certain number of copies corresponding to some Ct cutoff value, it is quite likely that of any three replicate sub-samples from a common sample with a variant concentration far from the LOD at least one will exhibit a Ct value greater than the cutoff and at least one smaller. Clearly at least one of these is just a “statistical fluctuation”. But is the statistical outlier the nominally positive replicate and the sample is thus “really negative” or just the opposite – the replicates are all “really positive” and the replicate with the high Ct value was merely a statistical fluctuation? Obviously this method of reporting the assay results does not provide a satisfactory measure of the statistical strength of the morph assay evidence.

Notes and References

1. *Review of the Scientific Approaches Used During the FBI’s Investigation of the Anthrax Letters*, National Research Council, (The National Academies Press, Washington DC, 2011).
2. S. Velsko, “Amerithrax ‘Morph’ Analysis: Statistical Issues”, Presentation to GAO September 18, 2011; LLNL-PRES-583052. A critical assumption in this analysis is that the RMR-1029 samples were true replicates in the sense that their production involved removing similar volumes of well-stirred material from the flask, and those volumes were diluted or cultured in the using the same technique. Thus the variance in replicate results represent the statistical performance of the assay protocols on identical samples. Unfortunately we have no description of the actual process used to generate the samples.
3. A compilation of the RMR-1029 replicate assay results is provided in: Ranajit Chakraborty, “Final Report of RFO-FBI-2008/AMERITHRAX Work (Statistical Analysis of Amerithrax Data)”; Appendix 2 page 81; Batch 2 Module 10.
4. The method for preparing these samples from the repository materials is not known to us at this time.
5. Applied Biosystems Application Note, “Real Time PCR: Understanding Ct”, Publication 136AP01-01, May, 2008.
6. “Assay Development Morph E Validation”, Appendix A, pages 59-81; Genetic

Mutation E (TIGR), Batch 2, Module 9.

7. Taylor LR, "Aggregation, variance and the mean", *Nature* 1961; **189**: 732–735.

8. S. Velsko, "When is a trace detection 'significant'?", LLNL-PRES-637099; BioDetection Technologies 2013, Alexandria Virginia, June 19, 2013.

9. Genetic Mutations B and D; Assay Validity Report Task Order Number 2: Morph D Deletion; Indexed Raw Data Outputs; pages 413-661; Batch 2 Module 7. If a direct comparison with the TIGR data is desired, the IITRI data can be converted to copy number by the simple expedient of using the conversion factor of 1×10^6 genomic copies/ng provided by CBI (Batch 2 Module 5, Page 578.)

10. Assay Development Morphs A1 – A3 Final Report (Including Validation), Validation Study of Morph A1 Protocol p 649-711 and Validation Study of Morph A3 Protocol p 868-881; Molecular Analysis of Pathogen Strains and Isolates and Genetic Mutations A1 and A3; Batch 2 Module 5

11. Genetic Mutations B and D - MRI data; Batch 2 Module 8.

12. Wayne S. Kendal, "A probabilistic model for the variance to mean power law in ecology", *Ecological Modelling* 1995; **80**:293-297.

13. The investigative summary released by the Department of Justice states: "the scientific analysis coordinated by the FBI Laboratory determined that RMR-1029, a spore-batch created and maintained at USAMRIID by Dr. Ivins, was the parent material for the anthrax used in the mailings." Note that the term "parent material" is ambiguous since it leaves open the possibility that a stock derived from RMR-1029 rather than RMR-1029 itself could have been the proximate source of the mailed material. It is, however correct to say that RMR-1029 cannot be excluded as a source of the mailed *B. anthracis*. *Amerithrax Investigative Summary*, available at <http://www.justice.gov/amerithrax/docs/amx-investigative-summary.pdf>