UCRL-TR-220803

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Bioagent Sample Matching using Elemental Composition Data: an Approach to Validation

S. P. Velsko

April 24, 2006

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 the University of California nor any of their employees, makes any warranty, express 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 the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Bioagent sample matching using elemental composition data: an approach to validation

Stephan P. Velsko
Lawrence Livermore National Laboratory
April 21, 2006

Abstract

Sample matching is a fundamental capability that can have high probative value in a forensic context *if proper validation studies are performed*. In this report we discuss the potential utility of using the elemental composition of two bioagent samples to decide if they were produced in the same batch, or by the same process. Using guidance from the recent NRC study of bullet lead analysis and other sources, we develop a basic likelihood ratio framework for evaluating the evidentiary weight of elemental analysis data for sample matching. We define an objective metric for comparing two samples, and propose a method for constructing an unbiased population of test samples. We illustrate the basic methodology with some existing data on dry *Bacillus thuringiensis* preparations, and outline a comprehensive plan for experimental validation of this approach.

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express 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 the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

1. Introduction

Matching samples of materials found at a crime scene with those associated with a person accused of a crime is a well-established method in forensic science. However, the probative power and admissibility of such evidence in the courtroom is receiving greater scrutiny in recent years due, in part, to the increasingly rigorous interpretation of the Federal Rules of Evidence as represented in the Daubert[1] and Kumho[2] decisions, and partly due to the great success of DNA evidence, which is often cited as an exemplar[3-6]. The importance of this trend is reflected in the recent NRC study of bullet lead analysis, and the subsequent withdrawal of this method by the FBI[7-9]. The fundamental scientific basis for the validity of even long-accepted "classical" forensic practices such as fingerprint and hair analysis have been recently questioned[10]. It is important to note that in most of these cases, the actual analytical method used to characterize the samples is not in question. The object of criticism is more often the scientific basis for assigning statistically meaningful inferential power to such data.

In microbial forensics investigations, it is often desirable to determine if an agent used in a crime or act of terrorism can be matched to the agent used in another, similar event, or to material recovered from a laboratory suspected too be the source of the agent[11]. The potential probative value of sample matching in a microbial forensic context can be illustrated by two scenarios:

(1) Envelopes containing *B. anthracis* powders are received at two or more separate locations, on significantly different dates, and the envelopes themselves do not look similar. A finding that the agent came from a common source greatly increases the odds that the same person or group is responsible for both events.

(2) Material from an envelope is compared with material from a stock found in a suspect source laboratory. A match between the characteristics of the materials clearly increases the probable association between the two anthrax samples.

Genetic sequence information can be used to relate the organism in an agent to organisms found at suspected sources[12]. However, as was evident in the investigation of the Anthrax letters incident of 2001, several laboratories may possess isolates of genetically identical organisms[13]. In some cases the pathogen may be present in common in environmental settings and thus available to many, in principle. If the agent is a toxin such as ricin there may be insufficient residual genetic material to analyze. Thus, there are many reasons why additional means for relating bioagent samples are needed. It is natural to turn to chemical or physical analysis for a solution to this problem. While the courts have already considered pathogen DNA sequence evidence in a microbial forensics context[14], it has not yet deliberated on the use of chemical and physical evidence in a bioterror or biocriminal trial. However, we can be certain that if such an occasion arises, the scientific support for this kind of evidence will receive rigorous scrutiny with regard to its admissibility and evidential weight. Thus, it is important to build a strong scientific basis for drawing inferences about sample matching when using chemical or physical characteristics of agent materials.

In the realm of chemical and physical analysis, there are many possible analytical methods by which two samples could be compared. Elemental composition, isotopic ratios, and microscopic morphology are methods that are used in many other forensic contexts, and have natural extensions to microbial forensics. Experience has shown that superficial examination of physical features of samples made by different methods may not be very revealing. For example, well-washed anthrax spore preparations tend to look similar, regardless of the growth medium and preparation method used. On the other hand, when samples are not thoroughly washed, the bulk morphology of two samples drawn from the same batch of bioagent can be quite different[15], presumably because of inhomogeneity introduced by the drying process. Thus, simple visual or microscopic inspection is not necessarily a reliable way to relate two samples.

Among the various methods that can be used to match bioagent samples, elemental analysis has several favorable features. First, it uses well-understood techniques that are familiar in other forensic contexts. For example, bulk elemental analyses by Inductively Coupled Plasma – Mass Spectrometry (ICP-MS) or Inductively Coupled Plasma – Optical Emission Spectroscopy (ICP-OES) have been used to characterize glass[16], office document paper[17], and foils[18]. In addition, there are well-developed techniques for non-destructive elemental analysis of bulk samples such as X-ray Fluorescence (XRF), or that are applicable to trace samples such as Particle Induced X-ray Emission (PIXE) and Secondary Ion Mass Spectrometry (SIMS). Several agent sterilization procedures such as irradiation and dry heating do not affect the elemental composition of a sample, and hence ordinary multipurpose instruments can be used to do the analysis.

The elemental composition of a bioagent is the end result of the entire process used to produce it. For an agent like *B. anthracis,* growth in a culture medium may be followed by separation, washing, and drying steps. Certain materials may also be added to the finished agent as part of the "weaponization" process. As illustrated in Figure 1, each of these steps has an influence on the overall elemental composition of the agent, either through addition or removal of certain elements.

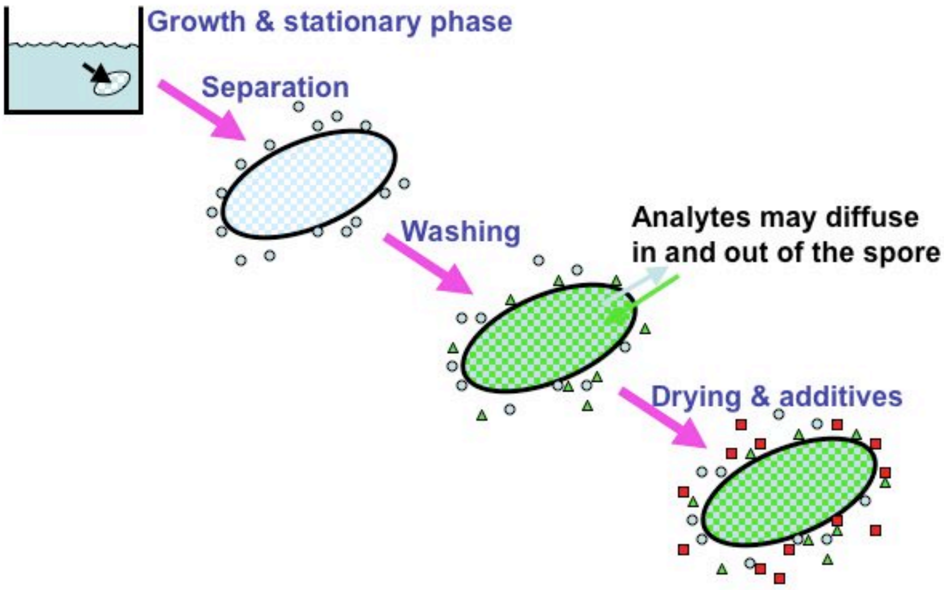
Growth & stationary phase
Separation
Washing
Analytes may diffuse in and out of the spore
Drying & additives

Figure 1. Schematic of steps in the growth and processing of a bioagent. Each step may add or remove elements to some degree.

Intuitively we expect that:

- The variance among elemental concentrations among samples from a single batch will be smallest, since each agent particle has experienced identical, or nearly identical growth and processing conditions;

- The variance among elemental concentrations from different batches made by the same process and the same materials may be somewhat larger due to uncontrolled random variations in how the process is carried out;

- Variance among elemental concentrations from batches made by the same nominal process but with different sources of starting materials will be larger still, due to variation in the elemental composition of starting materials from different sources; and

- The largest variance in elemental composition is expected among batches made by completely different processes utilizing different materials.

However, experimental studies are required to determine *quantitatively* and *statistically* the extent to which these intuitive ideas are correct when one considers a representative variety of growth and production processes. Ultimately, such studies should provide a sound statistical basis for deciding whether it is likely that two samples came from the same batch of material, or were made by different processes.

While it is clear that chemical or physical characteristics could provide information that pertains to agent preparation, we should be careful not to exaggerate current capabilities in this regard. A recent paper has claimed:

> *"... elemental signatures ... are useful for separating B. subtilis spores based on culture media, and the method may thus be applied to microbial source attribution in the future."*[19]

In fact, however, the reliability with which this or any other specific information about the growth and preparation procedure can be deduced from chemical and physical analysis remains to be demonstrated. Similarly, the application of such information to "source attribution" will only be possible with considerably more extensive validation than was presented in the cited publication. A careful evaluation of the use of elemental data for sample matching purposes is a critical first step in this direction.

Recent events in a related arena – the NRC analysis of bullet lead examinations and the subsequent discontinuation of this method by the FBI laboratories – provide an example of how careful scientific validation is *critical* if sample matching evidence is to be admissible and have defensible probative value. A new paradigm for conducting the evaluation and validation of forensic test methods is emerging, based on explicit likelihood analysis[20]. Under this paradigm, the key steps in evaluating and validating the application of elemental analysis to bioagent sample matching are:

Defining the "population" – i.e. the complete set of growth and preparation methods that might be used to generate the bioagent, and all of the potential sources for starting materials used in these processes.

Defining the signature – i.e. choosing the set of elements whose concentrations are to be determined in the method.

Defining an objective metric for decision – i.e. a (scalar) quantity defined in terms of the elemental concentrations that can be used to decide if two samples are related or not.

Characterizing the population – i.e. collecting elemental composition data on a set of bioagent materials generated by representative sampling of growth and preparation methods.

Evaluating the "receiver-operating characteristic (ROC)" – i.e. determining the dependence of the false positive and false negative rates of the method on the value of the chosen metric.

In this report, the term "validation" will refer to a two-phase process. First, the performance of the test is *evaluated* on a set of samples drawn from a representative population of samples. The test performance is expressed as a ROC curve. In the second phase, the performance is *validated* by a blind test on a completely independent set of samples drawn from the same population, or from a related population. This evaluation/validation paradigm is explicitly adapted from the literature on clinical and medical diagnostics. It should be noted that the definition of the sample "population" is one of the most critical steps in applying the evaluation/validation paradigm to the sample matching problem. It is essential that the sample sets used in this process are chosen in a defensibly unbiased way. This issue occupies a substantial part of the discussion in subsequent sections.

We will begin by outlining a method of analysis based on a standard framework that uses the receiver-operating characteristic (ROC) and likelihood ratios. An analysis of attribution that derives key quantities within this framework is contained in appendix 1. The application of this approach will then be demonstrated using recent data generated at Los Alamos National Laboratory based on samples generated at Lawrence Livermore National Laboratory. It should be emphasized that this analysis is merely presented for illustrating the method, because the data set is severely limited in extent. We then consider in more detail the nature of the "population" of growth and processing methods, and suggest a scheme for obtaining a representative set of samples for a more thorough evaluation and validation of the sample matching analysis method.

2. A likelihood ratio framework for bioagent sample matching analysis

In this section we describe a simple statistical framework for analyzing elemental data for purposes of drawing conclusions about the relatedness of samples. The terms "process", "batch", "replicate batch", "non-replicate batch" and "replicate sample" are used ubiquitously in this report and have the following definitions:

Process – a recipe following a fixed set of instructions for producing the finished agent. For bacteria and viruses, this includes both growth and post-growth processing. For biological toxins, it includes processes for extraction, purification and any subsequent chemical or physical treatment of the agent. Processes are distinguished by the types and amounts of materials used to make the agent, including growth media, chemical additives, solvents, etc.; and by choices of physical methods such as separation, drying and milling.

Batch – a single volume of material made from a single set of starting materials and following a fixed procedure (i.e. a "process", defined above.) For the case of a bacterial agent, a batch might be material produced from a single bench-top fermentation vessel, a set of shake flasks containing a common growth medium, or a set of agar plates poured from the same volume of prepared growth medium. In the latter two cases, material pooled from a set of shake flasks or agar plates and subsequently processed together would certainly constitute a single batch.

Replicate batches – A set of batches made by the same process where there is intentional duplication of materials and physical process parameters; material differences among replicate batches arise from un-intentional or uncontrolled changes in material quantities or parameters.

Non-replicate batches – A set of batches made by a common process, but where one or more factors are the result of independent choices among materials, equipment, or parameters not rigorously defined by the process description. An example would be where a bacterial culture is grown by the same method in two independent laboratories, where each lab obtains its materials from independent sources, or the staff uses slightly different procedures as interpreted from the same description of the process.

Replicate samples – samples drawn at random from a single batch of material, regardless of its homogeneity.

In general, bulk elemental analysis of two bioagent samples could result in one of three possible decisions:

- They were drawn from the same batch of original material. We will refer to this hypothesis as B_0

- They were made by the same process (but possibly non-replicate batches, using starting materials from different sources or different lots, or with slightly different process parameters). We will refer to this hypothesis as P_0.

- They were drawn from batches made by different processes. This is simply the negation of the same process hypothesis, $\overline{P}_0$, but also clearly implies $\overline{B}_0$.

There is also the possibility that the sample was made by pooling different batches, but in general it would be difficult to discern this by bulk analysis alone. Techniques that can analyze single agent particles, such as PIXE or SIMS might be more appropriate, but we will not treat this problem here.

An elemental analysis method like ICP-OES can measure the concentrations of a wide variety of elements down to ppb levels[21] (depending on the element and sample size.) The set of concentrations of those elements (e.g. Na, K, Ca, Sr, Mn, Zn, …) define a characteristic "elemental concentration vector" associated with a sample. The particular set of elements chosen for sample matching analysis must be determined, at least in part, by empirical considerations, such as which elements are highly likely to be above the detection limit for a given sample size for the chosen elemental analysis method. Given the elemental concentration vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ for samples 1 and 2, one can define a distance metric $\Delta_{12}(\mathbf{V}_1,\mathbf{V}_2)$ to describe the difference between the elemental signatures of the two samples. There are various ways that Δ_{12} can be defined, and we will choose a particular formula in the next section. In any case Δ_{12} is the fundamental statistic that is used to decide if two samples are more or less likely to be related.

The premise behind sample matching analysis is that Δ_{12} is materially relevant to the question of whether the two samples being compared came from the same source, i.e, from a common laboratory or a common batch of biological agent. In Appendix 1 the inferential power of an observation of Δ_{12} for two samples is described by the standard Bayesian formula

$$O(L_0|\Delta_{12}) = LR(\Delta_{12})\bullet O(L_0) \qquad (1)$$

Where L_0 is the hypothesis that the two samples originated in the same laboratory, and $O(L_0)$ and $O(L_0|\Delta_{12})$ respectively are the prior and posterior odds that L_0 is true. Thus, the materiality test[22,23] for the relevance of Δ_{12} in court is that $LR(\Delta_{12}) > 1$. An explicit analysis relating $LR(\Delta_{12})$ to the hypotheses L_0, B_0 and P_0 are given in Appendix 1.

The ultimate object of an evaluation/validation study is to show that the test procedure in question provides defensible (and useful) estimates of $LR(\Delta_{12})$. To achieve this, it is necessary to obtain a set of samples that represent an unbiased sampling of a "population" of laboratories, processes, batches, and replicate samples, determine their elemental composition, and empirically determine the distributions of Δ_{12} values for the sub-populations of sample pairs that are and are not from the same batch or process. The standard way to summarize this data is the receiver operating characteristic (ROC) curve[24], which displays the probability of true positive findings against the number of false positives as the test statistic increases or decreases in value. The value of LR can be estimated for any value of Δ_{12} from the ROC curve. This approach has been adopted in a number of contexts where critical decision-making is dependent on the results of experimental tests, including clinical testing and medical diagnosis[25-35].

Note that in this scheme, legal testimony about the significance of a particular value of Δ_{12} never needs to state that there is a "match" between samples. The relevant (and admissible) evidence is that the observed value of Δ_{12} increases (or decreases) by a certain amount the likelihood that the samples were drawn from the same batch or made by the same process. Another advantage of the ROC curve approach to characterizing sample matching analysis is that objective criteria then exist for comparing to other tests (using different analysis methods or different definitions of Δ_{12}) that have the same aim.

The sets of test samples used to construct and validate the ROC curve are drawn from a selection of M_L laboratories, each independently producing M_B replicate batches of agent for each of M_P distinct processes, where each batch is divided up into N replicate samples. The structure of the conceptual M_L x M_P x M_B source "population" is discussed in more detail in section 4. We suggest an explicit scheme for an evaluation/validation study involving different laboratories and processes. However, before describing this study, it is useful to apply the ROC analysis to some existing data to illustrate the method and to obtain some useful results to inform the plan.

3. Application to some existing data

The easiest way to illustrate the procedure outlined in the last section is to explicitly work through some existing data. The data set we shall use is a set of elemental concentrations for a set of *Bacillus thuringiensis israelensis* (*Bti*) samples that were grown and processed using a variety of methods. A summary of the preparation methods used for these samples is given in Table 1. The elemental concentrations were measured by Tom Yoshida at Los Alamos National Laboratory using ICP-OES and presented at the 3rd quarterly review meeting of the NBFAC National Laboratory R&D program on January 19th, 2006[36]. For convenience this data is presented in Table 2.

Several points must be noted about this data set. First, the sample matrix (Table 1) contains only part of the idealized $NxM_LxM_BxM_P$ set of samples discussed above. All of the samples were made in one laboratory. (In the language of Appendix 1, the prior probability of L_0 is unity.) Seven distinct processes are represented. Only one process has more than 1 batch associated with it, and there are no true replicate samples associated with any one batch. The absence of analytical replicates is an unfortunate consequence of limited sample availability. In the absence of a better estimator for variation among replicates we will use the data for acetone and lyophilized samples as if it were a replicate pair. This effectively reduces the number of distinct processes to 4. While this data set is far too small to provide accurate assessment of a ROC curve for the sample-matching test, it is large enough to illustrate the method.

Table 1. Description of Bti samples used to generate the data in Table 3.

Sample ID	Date of manufacture	Growth method	Growth medium	Washing method	Drying method
B1	06-01-04	Fermentor	G	SDS detergent + 2x water wash	Acetone
B2	06-01-04	Fermentor	G	SDS detergent + 2x water wash	Lyophilization
B3	06-21-04	Fermentor	G	SDS detergent + 2x water wash	Acetone
B4	06-21-04	Fermentor	G	SDS detergent + 2x water wash	Lyophilization
B7	09-20-04	Agar plate	G	2x water wash only	Lyophilization
F2	10-18-04	Agar plate	G	Cascade detergent + 2x water wash	Lyophilization
G2	10-18-04	Agar plate	G	Cascade detergent + 2x water wash	Acetone
H2	10-18-04	Agar plate	NB	Cascade detergent + 2x water wash	Lyophilization
I2	10-18-04	Agar plate	NB	Cascade detergent + 2x water wash	Acetone

Although Yoshida presented a longer list of elements, Table 2 displays only the 14 elements whose concentrations were the largest, and were detected in every sample. Some of these, like sodium, display little variation from process to process, while others appear to have larger variations. There is no reason to believe that the 14 elements chosen are the best or most efficient set for the sample-matching test.

Table 2. Elemental data from Yoshida (reference 36.)

Element	Concentrations (ppm)								
	B1	B2	B3	B4	B7	F2	G2	H2	I2
Na	4739	4057	3922	3356	5277	5542	3587	4625	5067
Rb	0.1	0.09	0.35	0.3	0.25	0.2	0.19	0.91	1.2
Mg	2844	6288	4510	5201	3166	3023	2990	3509	3600
Ca	75829	81136	62745	87248	29024	23687	25411	22329	24000
Ba	9.4	8.1	4.7	6.4	1	0.56	0.59	2.9	3
Sr	12	12	8.1	11	16	2.2	2.1	101	105
Al	2699	2403	1887	2037	1389	1266	1267	1481	1371
Mn	31357	24630	18477	21433	3429	1082	1147	41	19
Fe	3081	2840	1961	2685	237	126	90	80	80
Cu	325	334	352	168	186	214	153	12	10
Zn	3164	2825	2290	3195	347	178	171	65	79
Ti	506	418	361	738	196	194	194	164	182
Mo	1.5	1.1	0.88	0.84	0.4	0.3	0.24	0.21	0.21
Pb	20	19	4	4.4	0.17	0.68	0.55	0.06	0.06

Our analysis of this data is based on a pairwise distance estimator Δ_{12} defined by

$$\Delta_{12} = [4\bullet N^{-1}\bullet \sum(1/\sigma_l)^2(C_{1l} - C_{2l})^2/(C_{1l} + C_{2l})^2]^{1/2} \qquad (2)$$

where the sum over l is over the N elements in the fingerprint, and σ_l^2 is a statistical weighting factor discussed below. In this expression, normalization of the pairwise difference by the average of the pair helps to control for large changes in absolute concentration from element to element. When $C_{1l} >> C_{2l}$ or $C_{2l} >> C_{1l}$, the square of the difference divided by the sum approaches 1. Thus, the maximum value that Δ_{12} can attain is

$$\Delta_{max} = [4\bullet N^{-1}\bullet \sum(1/\sigma_l)^2]^{1/2} \qquad (3)$$

The first issue that needs to be addressed is how to define and estimate the weighting factor σ_l^2. For each element l, σ_l^2 is an estimate the variance of the quantity d_l defined by:

$$d_l = 2\bullet(C_{il} - C_{jl})/(C_{il} + C_{jl}) \qquad (4)$$

where i and j stand for independent pairs of data points drawn from the population of concentration values $\{C_{il}\}$ for that element determined by the replicate measurements on a batch from a particular process. (In the subsequent discussion, we will drop the index p since we will always be concerned with data from one particular process.)

The best way to estimate this variance is by making measurements on a number of replicate samples from one batch each of a set of representative processes. For each element l, this produces a set of concentration values $\{C_{pil}\}$ that represent the variance found in a single batch for each process p. There are two sources for this variation: (1) analytical variation due to instrument noise or uncontrolled random procedural variations, and (2) actual sample variation between replicates from the same batch. These two will typically be convoluted together.

If there is not much variation among the measurements for a particular element, simple linearized error propagation can be used to estimate the variance in d_l from the variances in the concentration values for that element.

$$\sigma_l^2 = \mathrm{Var}(d_l) \approx 2(1/C_l)^2 \bullet \mathrm{Var}(C_l) \qquad (5)$$

However, if the variance for an element is apparently large, then a direct estimate must be determined by sampling pairs of concentration values drawn from $\{C_{il}\}$.

It should be noted that implicit in this formulation is the assumption that the variations in elemental concentration among replicates is uncorrelated. Thus, a more rigorous treatment would include the possibility of correlation. As part of a more general validation exercise, replicate experimental concentration data should be tested to see if the variations in concentration among the elements are correlated or not.

Since the data set in Table 2 has at most 2 "replicates" each for 4 of the 5 batches, it is not possible to estimate σ_l^2 in the prescribed way. Instead, we have simply used the difference between the two "replicate" concentration values divided by their average, and then calculated the root mean square of these values over the 4 batches to generate a σ_l^2 value for use in formula (2). The results of these calculations are summarized in Table 3. Based on the σ_l^2 values in Table 3, the value of Δ_{max} is determined to be 11.8.

Table 3. Values used for estimating σ_l^2.

Element	$2\bullet\lvert(C_1 - C_2)/(C_1 + C_2)\rvert$				σ_l^2
	B1 & B2	B3 & B4	F2 & G2	H2 & I2	
Na	0.156	0.156	0.42	0.092	0.058
Rb	0.106	0.154	0.052	0.28	0.029
Mg	0.76	0.142	0.011	0.026	0.150
Ca	0.068	0.32	0.07	0.072	0.029
Ba	0.148	0.3	0.052	0.034	0.0289
Sr	0	0.3	0.046	0.038	0.0234
Al	0.116	0.076	0.00078	0.078	0.006
Mn	0.24	0.148	0.058	0.74	0.158
Fe	0.082	0.3	0.34	0	0.0531
Cu	0.028	0.7	0.32	0.182	0.157
Zn	0.114	0.32	0.04	0.194	0.0387
Ti	0.19	0.68	0	0.104	0.127
Mo	0.3	0.046	0.22	0	0.0351
Pb	0.052	0.096	0.22	0	0.0151

Finally we used equations (2) and (3) to calculate the normalized quantity

$$\Delta = \Delta_{12}/\Delta_{\max}$$

for each pair of samples in Table 2. This data is shown in Table 4.

The data in Table 4 is used to generate a ROC curve for batch-matching in the following manner. For each observed value of Δ, the number of sample pairs that had values less than or equal to Δ and were from the same batch (i.e. for which B_0 was true) is tabulated. Similarly, for each observed value of Δ, the number of sample pairs that had values less than or equal to Δ and were from *different* batches (i.e. for which B_0 was false) is tabulated. The cumulative numbers of "true positives" and "false positives" thus associated with each Δ are then respectively divided by the total number of true positive pairs (4) and true negative pairs (32) represented in this data set. By this process, each observed value of Δ has associated with it the fraction of true positives and fraction of false positives that would occur if we chose that value of Δ to be the criterion for deciding that the two samples were from the same batch. By plotting the true positive values against the true negative values we generate the ROC curve shown in Figure 2. The same procedure involving the hypothesis P_0 can be used to generate the ROC curve for process matching, shown in Figure 3.

Table 5. Normalized delta values (Δ) for pairwise sample comparisons.

	H2	I2	F2	G2	B1	B2	B3	B4	B7
H2	0	0.067	0.53	0.52	0.67	0.64	0.60	0.62	0.43
I2		0	0.53	0.52	0.68	0.67	0.61	0.63	0.43
F2			0	0.074	0.64	0.62	0.52	0.56	0.35
G2				0	0.64	0.63	0.53	0.57	0.35
B1					0	0.078	0.34	0.30	0.58
B2						0	0.32	0.29	0.56
B3							0	0.11	0.49
B4								0	0.52
B7									0

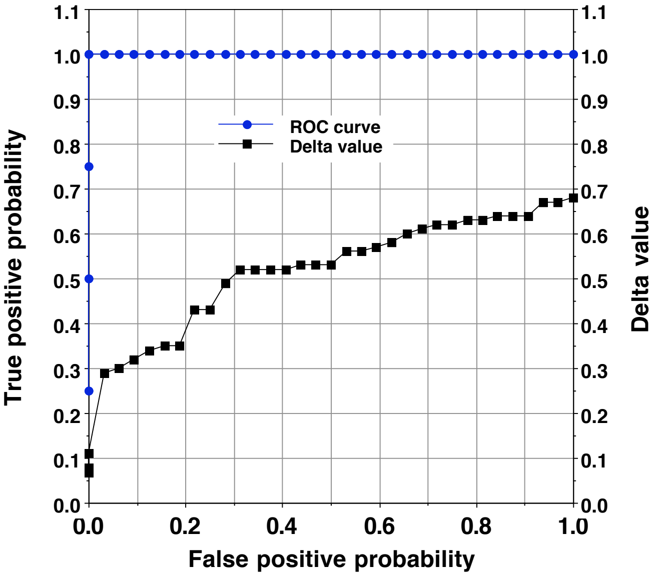

Figure 2, ROC curve for the batch-matching test generated from the data in Table 4. Also shown for convenience is the value of Δ associated with each pair of true and false positive fractions.

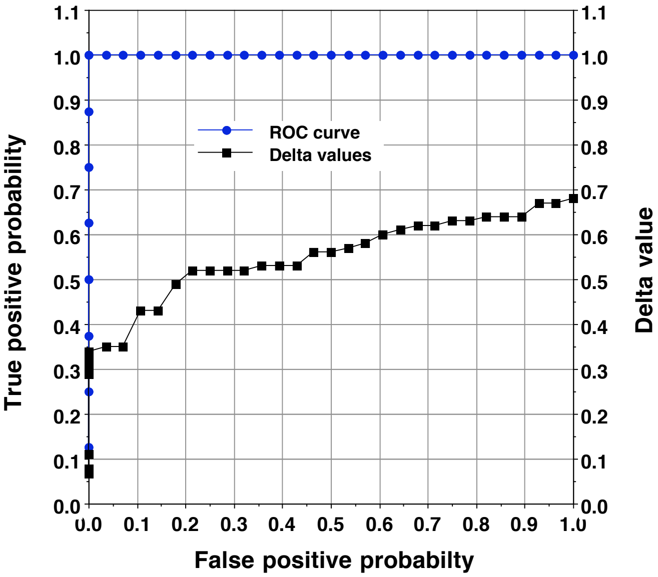

Figure 3. ROC curve for process matching generated from the data in Table 4.

It is clear from Table 4 that the values of Δ for pairs that came from the same batch are well separated from those that did not. For values of $\Delta \leq 0.11$ there are no false matches, and all true positives are "detected." This is clearly reflected in the corresponding ROC curve of Figure 2, which has the characteristic shape of a "perfect" test. Note that there is a relatively large gap between the largest value of Δ associated with B_0 (0.11) and the lowest value of Δ associated with $\overline{B}_0$ (0.29). In a larger, more diverse sample set it seems possible that this gap would be much smaller, or perhaps disappear entirely and cause the ROC curve to take a more typical ("imperfect") form.

Note, on the other hand, that the data in Table 4 does not support quite as clear a separation of the Δ values for samples generated by different production methods. The Δ values for (B7,F2) and (B7,G2), which represent pairs that were not made by the same method, are very close to the Δ value for (B1,B3) which were. While $\Delta \leq 0.34$ provides a clean separation between samples made by the same method, a larger sample population would almost certainly exhibit greater commingling of values for samples made by the same method and values of samples made by similar, but not identical methods. In that case *any* choice of Δ might result in non-zero values for the false positive and false negative test probabilities.

An unfortunate consequence of a ROC curve for perfect classification is that the likelihood ratio is effectively infinite when Δ is less than the threshold value and zero when Δ is greater than the threshold value. (In general, LR(Δ) is the slope of the ROC curve.) However, the point estimates of true positive and false positive probabilities for a given threshold value of Δ that are generated by the ROC curve are of little utility without estimates of their uncertainty. A number of standard methods exist for estimating this uncertainty at a desired confidence level (usually taken to be 95%)[37,38]. As an example, in Tables 5 and 6 we have generated estimates of confidence intervals for the conditional probability matrix for $\Delta \leq 0.11$ and $\Delta \leq 0.34$ respectively. Tables 5a and 6a give the test results for classification into B_0 and P_0 respectively. Estimators of the false positive and negative rates are nominally zero, but this is clearly not an accurate way to characterize the sensitivity and selectivity of the test, given the small size of the sample set. Therefore, confidence intervals for the conditional probability estimates were generated by using a web-based Bayesian method[39], and are displayed in Tables 5b and 6b.

In spite of its limited size and diversity, we can draw several conclusions from the dataset analyzed in this section. First of all, it may be possible to obtain a very clean classification of samples into "same batch and "different batch" categories using the measure Δ_{12} defined in equation (2). The results are consistent with the intuitive expectation that the false negative rate would be higher than the false positive rate for declaring that two samples come from the same batch. Similarly, intuition also suggests that there would be a higher false positive rate for declaring two samples to have been produced by the same method, because not all methodological variants will affect the elemental composition. However, the small number of samples, limited number of batches and restricted range of processes contained in this data set preclude us from drawing more precise conclusions. Clearly a more extensive validation study is required before this method can be used in a forensic application. In the next section we will propose a plan for such a study.

Table 5a. Summary of test results for $\Delta \le 0.11$ and B_0.

Test results for $\Delta \le 0.11$	**Samples were made in the same batch** **(B_0 is true)**	**Samples were not made in the same batch** **(B_0 is false)**	**# of test results**
Positive test results	**N = 4**	**N = 0**	**N(+) = 4**
Negative test results	**N = 0**	**N = 32**	**N(-) = 32**
Number of tests	**$N(B_0) = 4$**	**$N(\bar{B}_0) = 32$**	**$N_{total} = 36$**

Table 5b. Table of estimated bounds on the conditional probabilities for the "same batch" test, derived from Table 5a and 95% Bayesian confidence limits.

	Samples were made in the same batch **(B_0 is true)**	**Samples were not made in the same batch** **(B_0 is false)**
Positive test result **($\Delta \le 0.11$)**	**$P(\Delta \le \Delta_b | B_0) \ge 0.55$** **True Positive**	**$P(\Delta \le \Delta_b | \bar{B}_0) \le 0.087$** **False Positive**
Negative test result **($\Delta > 0.11$)**	**$P(\Delta > \Delta_b | B_0) \le 0.45$** **False negative**	**$P(\Delta > \Delta_b | \bar{B}_0) \ge 0.913$** **True negative**

Table 6a. Summary of test results for $\Delta \le 0.34$ and P_0.

Test results for $\Delta \le 0.34$	**Samples were made by the same process** **(P_0 is true)**	**Samples were not made by the same process** **(P_0 is false)**	**# of test results**
Positive test result **($\Delta \le \Delta_p$)**	**N = 8**	**N = 0**	**N(+) = 8**
Negative test result **($\Delta > \Delta_p$)**	**N = 0**	**N = 28**	**N(-) = 28**
Number of tests	**$N(P_0) = 8$**	**$N(\bar{P}_0) = 28$**	**$N_{total} = 36$**

Table 6b. Table of estimated bounds on the conditional probabilities for the "same process" test, derived from Table 6a and 95% Bayesian confidence limits.

	Samples were made by the same process **(P_0 is true)**	**Samples were not made by the same process** **(P_0 is false)**
Positive test result **($\Delta \leq \Delta_p = 0.34$)**	**$P(\Delta \leq \Delta_p \| P_0) \geq 0.72$** **True Positive**	**$P(\Delta \leq \Delta_p \| \overline{P}_0) \leq 0.098$** **False Positive**
Negative test result **($\Delta > \Delta_p = 0.34$)**	**$P(\Delta > \Delta_p \| P_0) \leq 0.28$** **False negative**	**$P(\Delta > \Delta_p \| \overline{P}_0) \geq 0.902$** **True negative**

4. An experimental design for a sample matching validation study

While the results in section 3 suggest that a useful sample-matching test can be based on the elemental profile of a biological agent sample, the limited nature of the sample set leads to considerable uncertainty about the true ROC curve for such a test. This section discusses a potential method for determining better estimates of test performance. The fundamental idea is to construct a more representative population of agent samples (using a *B. anthracis* surrogate that is much closer than *B. thuringiensis*) and to use samples from this population in two distinct phases of testing. In the first phase, a set of samples are generated and analyzed to determine the ROC curve. In the second phase, samples are randomly selected from well-characterized archival materials whose provenance is known, or generated independently from the original set, and blindly evaluated to validate the prior performance estimate.

A. General considerations

Before discussing more particular features of the validation process, it is of great value to review the potential sources of false positive and negative rates in sample matching analysis of biological agents. Clearly, an unbiased design of the validation exercise must provide a fair chance that these causes of error are embodied in the experimental sample set. Table 7 summarizes the potential reasons for errors in matching bioagent samples. Two sources of error included in this list, data errors (i.e. mistakes in calculation or recording of data) and contamination can be controlled if the laboratories that undertake such measurements have strict QA/QC procedures in place. For purposes of this report, we will assume that these types of errors are improbable. In addition, we assume that the cited sources of error are far more significant in practice than any errors that might arise due to intrinsic uncertainty in the measurement of elemental concentrations. As pointed out in section 3, measurement uncertainty is folded in with the variance of elemental concentrations determined from analytical replicates from the same batch of material. Note also that we have included for completeness the theoretical possibility that growing organisms such as bacteria may have mechanisms that control the maximum or minimum concentrations of certain trace elements within the cell. Thus, in an extreme case, the concentrations of elements in an agent powder might be nearly independent of their concentrations in the starting materials. We know of no such case in practice, but it does lead to the precaution that the organism used in the validation experiments be physiologically similar to the threat agent for which it is a surrogate. The remaining errors will be the primary focus of our considerations.

The exercise with the *Bti* samples suggests some additional points about decision errors in sample matching. From our results in section 3 we anticipate that very few circumstances will make the signatures of two samples coming from different processes as similar as two samples drawn from the same batch. In other words it is not very probable that there will be a fortuitous match between the (multi-element) signatures of two samples grown by very different methods. However, samples made by certain growth medium formulations (e.g. G medium and Modified G medium, which differ only by the addition of glucose) may not be distinguishable, assuming that the medium components they have in common are drawn from the same lot. Certain process differences may not lead to significant differences in the elemental composition of the agent. For example, we know from the *Bti* samples that certain differences in drying procedure make very small changes in the elemental signatures. Powder grinding is also a potentially neutral step as far as elemental signatures are concerned. It may be necessary to restrict the scope of the test to a set of “distinguishable processes” and place certain variants into indistinguishable classes.

We may also anticipate that there are a fair number of reasons that two batches of agent made by nominally identical processes could give significantly different elemental profiles. Technicians are known to make unintentional alterations in medium preparation such as using poorly approximated mineral salt concentrations. Accidental contamination with extraneous materials could alter the elemental profile of a bio-agent preparation. Also, small unintentional differences in aeration or temperature conditions during growth could affect growth rate or sporulation time, and modify the uptake of trace elements by spores.

Ideally, all of these potential sources of error and unintentional variation should be reflected in the sample set used to determine the ROC curves of sample matching tests. In practice, of course, it is not possible to predict the base rate for these errors, so it is not even possible to estimate the number of samples that would have to be generated to guarantee that such variation is fairly represented.

Table 7. Possible reasons for false positive or false negative test results.

Test Result	Test	
	Same Batch	Same Process
False positive (Declaring two samples to be drawn from the same batch or made by the same process when they are not.)	• Tight control on the repeatability of medium lot, preparation procedure, and other process parameters • High uniformity in manufactured, premixed or pre-poured media • Extreme physiological "leveling" of trace element composition in microbe	• Insensitivity of elemental composition to certain process steps, e.g. drying, washing (e.g. samples B7, F2 and G2 from the *Bti* sample set.) • Similarity of certain unit processes, e.g. shake flask vs. fermenter. • Closeness of certain media formulations e.g. among *Bacillus* media in the same category (see Table xx). • Accidental, random identity in trace element composition of two different media • Moderate physiological "leveling" of trace element composition in microbe
False negative (Declaring two samples to be drawn from different batches or made by different processes, when they are not.)	• Inhomogeneity in a single batch + measurement variance. • Contamination • Errors in data analysis or data recording	• Variation in elemental composition of starting materials, from source to source • Variation in sample preparation by the preparer. • Contamination

B. The bioagent sample "population".

In forensic sample matching problems that involve the comparison of manufactured products like lead bullets, glass, or agricultural commodities, there is a real "background population" of materials, and the major concern is that databases or reference sample collections contain an adequate and up-to-date representation of this population. In contrast, biological agents are not regularly manufactured, and there is an extremely limited real background population of already-made samples[40]. Instead, the "population" is an imaginary construct, consisting of all possible ways that someone might go about making an agent. In section 2 we introduced the idea of a fictitious sample population consisting of M_L "laboratories", M_P possible processes, M_B replicate batches from each process, and N replicate samples drawn from each of the different batches. The problem is how to generate a set of real samples that adequately represents a statistically valid sample of this imaginary space of possibilities.

Within this space, not every process is equally likely. Information from a variety of sources indicates that there are a countable number of processes that have been developed for growing and processing pathogens, and only a handful of these are “popular” among microbiologists engaged in biodefense work. Knowledge of these recipes flows to other laboratories through publications and reports, or when personnel move from one laboratory to another. Analysis of existing information indicates that when criminals or terrorists produce bioagents they draw upon existing information for growth and processing ultimately derived from Western open literature sources. Thus, the prior probability of seeing certain recipes is higher than for others. In the following discussion, we will restrict our attention to *Bacillus anthracis* production, recognizing that different considerations may apply to other microbial agents.

A particular feature of *Bacillus* growth is that a wide variety of growth media can be used. In addition to refined, commercial media (e.g. peptones, tryptones, glucose) less refined, complex medium formulations (e.g. corn steep liquor, molasses) can be used. The former are typical of bench-top fermentations in research laboratories, but may also be used at pilot scale in sophisticated industrial-scale state sponsored biological weapons programs, because it makes separation and purification of the agent easier. Complex medium formulations are often used for manufacturing *Bacillus thuringiensis* (*Bt*) based insecticides, and could be readily adopted for B. *anthracis* production, especially in cases where a *Bt* plant is surreptitiously being used. Table 8 summarizes this space of possibilities.

Table 8. Coarse-grained classification of *B. anthracis* growth medium formulations

	Bench-top	Large scale
Simple media	Most likely as terrorist or criminal activity	Sophisticated pilot or industrial-scale state programs
Complex media	Less likely as terrorist or criminal activity	Surreptitious conversion of industrial scale processes

It can be argued that the most likely scenario for terrorist or criminal production of *B. anthracis* involves bench-top methods using commercial refined media that would typically be found in the research laboratories in which the perpetrators received their training, or mentioned in the literature that they can readily access. Therefore, our discussion will focus on this situation.

To better define the population of benchtop scale *B.anthracis* production methods from which to draw samples for a validation study, we have assembled information from several sources: reports from threat assessment programs sponsored by the U.S. government, other reports, and open scientific literature. A table describing the composition of 23 distinct growth medium formulations used to grow *B. anthracis* is contained in the Excel file **MediumFormulation.xls** that accompanies this report. Table 9 contains a summary of this data. The media can be roughly classified into 7 distinct classes according to the major nutrient sources they contain. There were occasionally differences among the exact formulations quoted for some media, which is another potential source of variability in actual medium composition in practice.

Table 9. Summary of the 23 medium formulations found for B. anthracis growth.

Designator	Medium name	Medium class	# of citations
BHI	Brain-Heart Infusion	Beef Heart Infusion	12
LB	Luria-Bertani	Tryptone/Yeast Extract	10
BAgar	Bacto Blood Agar	Beef Heart Infusion	7
NSM	New Sporulation Medium	Tryptone/Yeast Extract	6
TSB	Tryptic Soy Broth	Tryptone/Soy	5
NB	Nutrient Broth	Peptone/Beef Extract	5
G	G Medium	Yeast Extract/Ammonium	4
R	R Medium	Chemically Defined	4
ModG	Modified G Medium	Yeast Extract/Ammonium	3
NBY	Nutrient Broth-Yeast Extract	Peptone/Beef Extract	3
LD	Leighton-Doi	Peptone/Beef Extract	3
SSM	Schaeffer's Sporulation Medium	Peptone/Beef Extract	3
MFA	Modified FA Medium	Tryptone/Yeast Extract	2
CADM	Casein Acid Digest Medium	Casein/Yeast Extract	2
CDSM	Chemically Defined Sporulation Medium	Chemically Defined	1
Liu	Liu Medium	Yeast Extract/Ammonium	1
MM	Miller-McBride Medium	Tryptone/Yeast Extract	1
PA	Protective Antigen Medium	Tryptone/Yeast Extract	1
NBA	FAO Nutrient Broth Agar	Peptone/Beef Extract	1
NSMP	NSMP-Modified	Casamino acid media	1
CDAM	Casein Digest Agar Medium	Casein/Yeast Extract	1
ATCC573	ATCC Bacillus Medium 573	Yeast Extract/Ammonium	1
ATCC552	ATCC Bacillus Medium 552	Peptone/Beef Extract	1

Furthermore, we have estimated rough measures of how "popular" a given medium is by counting the number of times the use of each medium formulation has been cited by independent laboratories. This was accomplished by searching PubMed for papers referencing "*B. anthracis*" in conjunction with a number of other terms, such as "vaccine", "sporulation", "expression", and "detection" and choosing the citations with freely available pdf files. These papers were then searched manually for any information on growth media used to culture *B. anthracis* or near neighbor surrogates including vaccine strains, *B. cereus*, and *B. thuringiensis*. Multiple references to a particular growth medium by the same research group were only counted as a single reference[41]. In addition, some general references on growth media and some specialty publications were consulted. The results in Figure 4 are the number of independent references found for each type of growth medium, a total of 78 references being found in this "random" search. Table 9 provides the key to the medium designators used in Figure 4.

The first 8 medium formulations (BHI, LB, BAgar, NSM, TSB, NB, G, and R) encompass 68% of the citations, while four additional formulations (ModG. NBY, LD, and SSM) encompass an additional 15%. The remaining 11 formulations account for only 17% of the citations, and appear to be seldom used within the microbiological community.

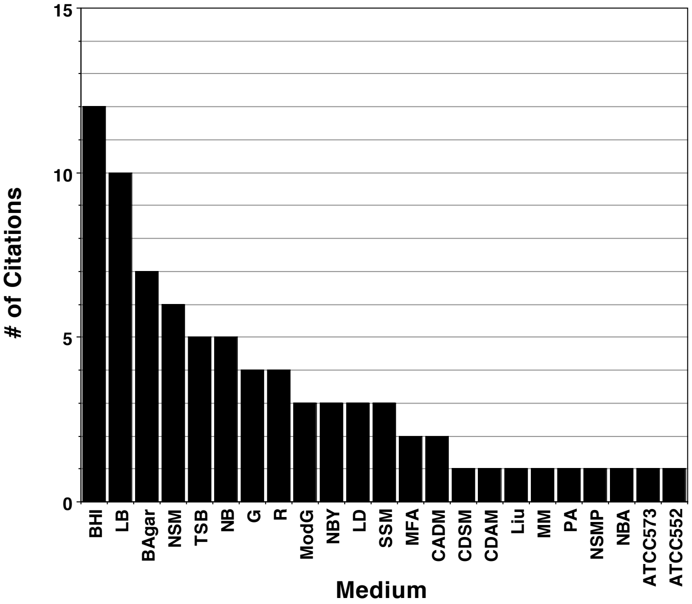

Figure 4. Distribution of 78 citations of various growth media used to grow *B. anthracis* or very near neighbors.

A variety of post-growth processes have been used for separating, washing, and drying *Bacillus* preparations subsequent to culture. In order to assess the “popularity” of each unit process step we surveyed a variety of reports and consulted with a limited number of experts in this area. Table 10 contains rough estimates of the statistical weight associated with various unit process steps derived from our survey. This table is a greatly simplified representation of the process choices that are available. Several more sophisticated methods for separation have been left out because they do not appear to be widely used at the bench-top scale. In addition, certain steps, such as detergent washing, encompass a variety of choices that are not captured here. Similarly, the weights assigned to each step are tendered with a great deal of uncertainty based on the limited sampling we were able to accomplish. Nonetheless, we believe that Table 10 is sufficiently representative to demonstrate the approach, and provides a basis for future refinement of this data.

A crude representation of the prior statistical weight of any given preparation method is simply the product of the weights associated with the choice of growth medium, growth method, separation method, washing and drying steps. This would be accurate if such choices were, in fact, independent. However, in reality end-to-end processes are usually treated as a whole, and thus it could be expected that many choices of individual process steps would be highly correlated. However, in the absence of a better base of information, the assumption of independence at least has the virtue of increasing the diversity of the validating sample set.

Table 10. Unit process steps and estimated statistical weights for *Bacillus* preparations.

Growth method	Separation	Detergent wash	Water wash	Drying
Agar plate 0.4	Centrifuge 0.75	No 0.75	2x 0.25	Lyophilize 0.75
Shake flask 0.4	Flocculation 0.25	Yes 0.25	5x 0.75	Acetone 0.25
Fermentor 0.2	-	-	-	-

Combining the information in Table 9 with that in Table 10 and ordering the product probability associated with each process from largest to smallest, one can arrive at a list that crudely represents the weighted population of processes for producing *B. anthracis* powders. In the present case, we have constructed such a list omitting the drying process, since this is known to have a very small affect on elemental composition. In addition, this list was edited to remove some entries that, for certain reasons, would be much less likely than the simple product of probabilities would imply. For example, the medium blood agar (BAgar) is clearly only associated with growth on agar plates, and not with shake flask or fermentor growth. Similarly the use of flocculants to separate spores from spent medium is exclusively used when growth is in liquid culture, not agar plates. With these amendments, the first 30 entries of this list are given in Table 11. It should be noted that fermentor growth appears only twice in this list as a consequence of its low probability relative to agar plate and shake flask growth.

The full list of processes is provided in an Excel spreadsheet **Processes.xls** that accompanies this report. This ordered list is a basis for selecting processes to generate samples used to validate the sample matching method. The statistical weight (product of probabilities) of each method is used to change the relative probability of selecting it at random from the entire “population” represented by the list. While the list in **Processes.xls** may leave out some plausible processes, it can be argued that it covers the most likely methods that would be used to produce *B. anthracis* preparations and represents a reasonable estimate of their relative likelihood.

We have used the built-in random number generation tool provided by Excel to generate a list of randomly sampled processes. The “Discrete” distribution type was selected, with the product probabilities associated with each process used to weight the selection. Table 12 contains the first 10 randomly selected processes.

Table 11. The 30 "most likely" process choices for *Bacillus* production. Key: A – Agar plates; S – shake flask; F – Fermentor; C – centrifuge; P – flocculation; N – no detergent; Y- detergent used; 5 – wash 5x with water; 2 – wash 2x with water.

Medium	Method	Separation	Detergent	Wash
BHI	A	C	N	5
BHI	S	C	N	5
LB	A	C	N	5
LB	S	C	N	5
Bagar	A	C	N	5
BHI	F	C	N	5
NSM	A	C	N	5
NSM	S	C	N	5
BHI	A	C	N	2
BHI	S	C	N	2
TSB	A	C	N	5
TSB	S	C	N	5
NB	A	C	N	5
NB	S	C	N	5
LB	F	C	N	5
LB	A	C	N	2
LB	S	C	N	2
BHI	A	C	D	5
BHI	S	C	D	5
BHI	S	P	N	5
G	A	C	N	5
G	S	C	N	5
R	A	C	N	5
R	S	C	N	5
LB	A	C	D	5
LB	S	C	D	5
LB	S	P	N	5
ModG	A	C	N	5
ModG	S	C	N	5
NBY	A	C	N	5

Table 12. The first 10 randomly selected production processes (Key same as in Table qq)

Process #[a]	Medium	Method	Separation	Detergent	Wash
1	BHI	A	C	N	5
89	ModG	A	C	N	2
21	G	A	C	N	5
96	SSM	S	C	N	2
5	Bagar	A	C	N	5
53	NSM	A	C	D	5
48	LB	F	C	N	2
74	LB	F	P	N	5
29	ModG	S	C	N	5
16	LB	A	C	N	2

[a] From the full ordered list in Processes.xls.

C. Constructing the set of test samples

The connection between sample matching and attribution described in Appendix 1 implies that there are particular “sub-populations” of pairs of test samples that must be included in any validation study. These “sub-populations correspond to hypotheses regarding the possible origin of the sample pairs and are summarized in Table 13.

Table 13. Partitioning of the test sample space derived from Appendix 1.

Test sub-population	Hypothesis from Appendix 1	Comment
Pairs of samples drawn from the same batch of material produced in a given laboratory	B_0L_0	The union of these two populations represent pairs of samples drawn from batches of material made in the same laboratory. $\{B_0L_0\} \cup \{\overline{B}_0L_0\} \equiv \{L_0\}$
Pairs of samples drawn from different batches of material produced in the same laboratory	$\overline{B}_0L_0$	
Pairs of samples drawn from two different batches of material made in two different laboratories	$\overline{B}_0\overline{L}_0 \equiv \overline{L}_0$	The set of samples drawn from the same batch produced in different laboratories is the null set. $B_0\overline{L}_0 \equiv \varnothing$
Pairs of samples drawn from batches made by the same process in the same laboratory	P_0L_0	The union of these two populations represent pairs of samples drawn from batches of material made by the same process
Pairs of samples drawn from batches made by the same process in different laboratories	$P_0\overline{L}_0$	
Pairs of samples drawn from batches made by two different processes in the same laboratory	$\overline{P}_0L_0$	The union of these two populations represent pairs of samples drawn from batches of material made by different processes
Pairs of samples drawn from batches made by two different processes in different laboratories	$\overline{P}_0\overline{L}_0$	

A sample set that captures these sub-populations in an unbiased way can be constructed by the following procedure: First, three or more processes are selected at random using the weighted “population” of processes described in the previous section. Secondly, three or more independent laboratories are selected to produce batches of material using the selected processes. A partial factorial design that reduces the total number of batches produced, but retains symmetry is shown in Table 14.

Table 14. A 3 x 3 partial factorial design for sample production.

	Lab 1	Lab 2	Lab 3	# batches per process
Process 1	Batch 1 Batch 2 Batch 3	Batch 5 Batch 6	Batch4	6
Process 2	Batch 4	Batch 1 Batch 2 Batch 3	Batch 5 Batch 6	6
Process 3	Batch 5 Batch 6	Batch 4	Batch 1 Batch 2 Batch 3	6
# batches per lab	6	6	6	Total # of batches = 18

Ideally, the laboratories would be selected at random from a larger pool of potential participants. Identical descriptions of the chosen processes are communicated to each lab, but the laboratories are conceived of as autonomous entities making independent choices of vendors and particulars of execution. A conservative bias could be introduced by suggesting to each lab that the similarity of the replicate batches is of high value to the exercise. It is assumed that, faced with the task of producing identical batches, a laboratory worker would plan to purchase sufficiently large lots of medium or numbers of pre-poured agar plates to complete the task when these are commercially available. Finally, it should be noted that the laboratories need not be physically distinct, but could conceivably involve different persons working in a common lab.

From each of the 18 batches, three replicate samples could be drawn for elemental analysis, leading to 54 separate elemental “vectors.” This is the basic block of data that underpins the evaluation/validation procedure outlined in the next section. Appendix 2 provides an explicit accounting of the sizes of the various sub-populations of pairwise comparisons generated by this data set.

D. Validating the ROC curve(s)

Appendix 1 demonstrates that there are two different ROC curves that are necessary to characterize the evidentiary power of a batch-matching test. One of these plots involves $P(\Delta \leq \Delta_b|B_0L_0)$ *versus* $P(\Delta \leq \Delta_b|\overline{B}_0\overline{L}_0)$ while the other involves $P(\Delta \leq \Delta_b|\overline{B}_0L_0)$ *versus* $P(\Delta \leq \Delta_b|\overline{B}_0\overline{L}_0)$. The data block generated by the procedure outlined above ought to have sufficient size to provide a reasonable representation of each curve. The process of validation then consists of independently generating a set of samples from the same population (in this case the population of bench-top *B. anthracis* production processes), performing elemental analysis on them, and comparing the ROC curves from that set to the previous curves. In the validation phase, the analyst should be “blinded” with respect to the origin of the samples, and the elemental analysis results should be reported in a coded fashion. The analysis of this data can be done in two ways. The first way is to use the existing ROC curve to make decisions about the validation set based on a chosen threshold value. Then the number of false positives can be compared to the number predicted by the ROC curve. The second way is to construct a ROC curve from the validation data itself, which is the preferred approach if there is sufficient data.

There are several approaches to comparing ROC curves[35,42]. One method is to fit the curves to a parametric form and then compare the parameter values to see if they are statistically different. Another method involves comparing the area under the ROC curves, which is a measure of the discrimination power of the test. At any value of the cutoff parameter Δ_b we can also test to see if the new probabilities are significantly different from the old by using a chi-squared test. In any case, the validation process can be iterated by using the validation set to “update” the ROC curve, and then testing the new curve against a new (independent) data set. The expectation is that the empirical ROC curve would converge on a form that best represents the performance of the test on the target population.

There are several potential ways to generate the set of samples used in the validation phase. One way would be to use existing archival samples that may be available from other laboratories. Another way is to use an independent block design similar to the one described in section 4C, where laboratories and processes are again randomly chosen from the available “population.” Alternatively, this can be done by using a series of several smaller block designs, e.g 2x2 versions of the 3x3 scheme in 4C. Ideally, the validation set would be similar in size to the set used to initially evaluate the ROC curves.

5. Concluding remarks

The intention of this report is to outline an approach to evaluate and validate sample-matching tests using elemental analysis as an example. An important feature of the chosen approach is the adoption of a framework that does not involve defining a “match”, but relies instead on estimating a likelihood ratio for any value of a defined pairwise comparison metric determined by elemental analysis. This philosophy is consistent with modern concepts of trace evidence analysis and with recommendations emanating from recent National Research Council studies on several forensic science issues.

There are numerous ways that the specific steps described here could be modified and possibly improved. For example. it is important to note that several segments of the larger “population” of *B. anthracis* production methods were left out of the treatment in this report. It is likely that these population segments would exhibit somewhat different ROC curves due to the increased variability associated with complex medium components and the increased chance for inhomogeneity among samples from large-scale production runs. In addition, even the set of bench-top processes considered here deliberately excludes a number of less likely processes. More extensive polling of both the literature and active research groups might provide a more accurate picture of the "population" of *B. anthracis* processes. It is also important to emphasize that the production of other agents, such as *F. tularensis* or *Y. pestis*, involve different types of growth media and post-growth processes.

It may be possible to construct improved classifiers that work better than Δ_{12} as defined above. For example, principal components analysis (PCA[43]) could be applied to the elemental concentration vectors to produce a variant of Δ_{12}. Given vectors of elemental concentrations on a reference set of agent samples prepared by a variety of methods, the first principle component is a linear combination of the concentrations of the elements that exhibits the most variation over the set of samples. Let C_{1n} and C_{1m} be the first principle components of the concentration data for two samples n and m. A metric for judging how similar the two samples are is:

$$\Delta_{mn} = 2(C_{1n} - C_{1m})/(C_{1n} + C_{1m})$$

By determining Δ_{mn} for pairs of samples from the same batch, replicate batches, non-replicate batches and from different processes, it should be possible to select a decision criterion (i.e. $\Delta_{mn} \leq \Delta_0$) for declaring that two samples are drawn from the same batch or were produced by different methods. Since this procedure uses the multi-element signature that shows maximum variation over the "population" of samples, the decision criteria so determined are conservative quantities.

Principle component analysis can also be used to decide if certain subsets of elemental concentrations are better than others for capturing the variability of elemental composition among agent samples. This is useful if it is desirable to reduce the number of elements that must be analyzed, or if one is comparing the merits of two different analytical methods that measure different sets of elements. A description of this procedure as applied to the bullet lead analysis problem is provided in reference 7.

The sample set that would be generated by the plan outlined in section 4 would also be useful for validating the use of other types of measurements for sample matching. For example, isotopic signatures could easily be evaluated by the same kind of approach, assuming a suitable metric analogous to Δ could be formulated. Approaches based on the morphology of samples may also be feasible if an objective metric could be formulated. Finally, the sample set outlined above could also contribute to the evaluation and validation of other SOPs e.g. analyses that aim to detect residual agar.

References

1. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

2. Kumho Tire v. Charmichael, 526 U.S. 137 (1999).

3. M.J. Saks and J.J. Koehler, "The Coming Paradigm Shift in Forensic Identification Science", Science **309**, pp.892-895, (2005).

4. D. Kennedy and R.A. Merrill, "Assessing Forensic Science", Issues in Science and Technology, Fall 2003, pp.33-34.

5. A.P.A. Broeders, "Of fingerprints, scent dogs, cot deaths, and cognitive contamination - a brief look at the present state of play in the forensic arena", Forensic Science International, (2005) Available on-line at www.sciencedirect.com.

6. P.C. Giannelli, "Forensic Science", Journal of Law, Medicine and Ethics, Fall 2005, pp.535-544.

7. National Research Council, Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison, Forensic Analysis: Weighing Bullet Lead Evidence, (National Academies Press, Washington DC, 2004).

8. M.O. Finkelstein and B. Levin, "Compositional Analysis of Bullet Lead as Forensic Evidence", Journal of Law and Policy **13**(1), pp.119-142, (2005).

9. "FBI Laboratory Announces Discontinuation of Bullet Lead Examinations", Press Release, Federal Bureau of Investigation, Washington, DC, September 1, 2005.

10. S.L. Zabell, "Fingerprint Evidence", Journal of Law and Policy **13**(1), pp.143-179, (2005).

11. B. Budowle, et. al., "Microbial Forensics: the next forensic challenge", International Journal of Legal Medicine, 119, pp.317-330, (2005)

12. B. Budowle, et. al., "Genetic analysis and attribution of microbial forensics evidence", Crit. Rev. Microbiol. **31**, pp.233-254, (2005).

13. Read, T.D. et. al., "Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*", Science 296, pp.2028-33, (2002).

14. State of Louisiana v. Richard J. Schmidt, Westlaw 699 So.2d 488.

15. S. Velsko, "Physical and Analytical Chemical Analysis: A key component of Bioforensics", UCRL-CONF-209735, Lawrence Livermore National Laboratory, 2005.

16. Montero, S. et. al., "Elemental analysis of glass fragments by ICP-MS as evidence of association: Analysis of a case", Journal of Forensic Sciences **48**, pp.1101-1107, (2003).

17. Spence LD, et.al., "Comparison of the elemental composition of office document paper: Evidence in a homocide case", Journal of Forensic Sciences **47**, pp.648-651, (2002).

18. Koons, R.D. et. al., "Comparison of household aluminum foils using elemental composition by inductively coupled plasma-atomic emission spectrometry", Journal of Forensic Sciences **38**, pp.302-315, (1993).

19. Cliff, J.B. et. al., "Differentiation of Spores of Bacillus Subtilis Grown on Different Media by Elemental Characterization Using Time-of-Flight Secondary Ion Mass Spectroscopy", Applied and Environmental Microbiology **71**, pp 6524-6530, (2005).

20. a. Lucy, D. *Introduction to Statistics for Forensic Scientists*, (John Wiley & Sons Ltd., West Sussex, UK, 2005); b. C. Aitken and F. Taroni, Statistics and the evaluation of Evidence for Forensic Scientists, 2nd ed., (John Wiley & Sons Ltd., 2004).

21. Strictly speaking, the analytical method *estimates* the concentrations. In this report, we will always assume that one concentration estimate per element is produced for each sample. That is, either the entire sample is consumed in each analysis or the results of multiple analytical replicates is averaged to produce a single concentration estimate. This assumption can clearly be relaxed at the cost of some extra complication in notation.

22. Imwinkelried, E.J., *The Methods of Attacking Scientific Evidence*, 4th Ed. (LexisNexis, 2004).

23. R. Lempert, "Modeling Relevence", Michigan Law Review Vol. 75, pp.1021-1057, (1977)

24. This is also referred to as the receiver-*operator* characteristic. This report uses the terminology most often found in the clinical diagnostics literature.

25. Zweig, M.H., "Evaluation of the Clinical Accuracy of Laboratory Tests", Arch. Pathol. Lab. Med., **112**, pp.383-386, (1988).

26. Spiehler, V.R., et. al., "Confirmation and Certainty in Toxicology Screening", Clin. Chem. **34**, pp.1535-1539, (1988)

27. Zweig, M.H., "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine", Clin. Chem. **39**, pp.561-577, (1993).

28. Spiehler, V., "Enzyme immunoassay validation for qualitative detection of cocaine in sweat", Clin. Chem. **42**, pp. 34-38, (1996).

29. Gryzbowski, M., "Statistical Methodology III. Receiver Operating Characteristic (ROC) Curves", Academic Emergency Medicine **4**, pp.818-826, (1997).

30. Zhou, K.H., et. al., "Smooth Non-parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests", Statistics in Medicine, **16**, pp.2143-2156, (1997).

31. van Erkel, A.R. and Pattynama, "Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology", European Journal of Radiology **27**, pp.88-94, (1998).

32. M. Greiner et. al., "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests", Preventive Veterinary Medicine **45**, pp.23-41 (2000).

33. Zhou, K.H., et. al., "Statistical Validation Based on Parametric Receiver Operating Characteristic Analysis of Continuous Classification Data", Academic Radiology, 10, pp.1359-1368, (2003).

34. Tom Fawcett, "ROC Graphs:Notes and Practical Considerations for Researchers", Kluwer Academic Publishers, 2004.

35. M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford Statistical Science Series 28; Oxford University Press, New York, 2003)

36. T. Yoshida, "Chem/Phys Signatures Elemental Analysis, 3rd Technical NL/NBFAC Quarterly Review, January 19th, 2006.

37. Newcombe, R.G., "Two-sided Confidence intervals for the sngle proportion: comparison of seven methods", Statistics in Medicine, **17**, pp.857-872, (1998).

38. T.D. Ross, "Accurate confidence intervals for binomial proportion and Poisson rate estimation", Computers in Biology and Medicine **33**, pp.509-531, (2003).

39. A number of programs for calculating exact confidence intervals for proportions are available on the Web. This calculation used a Bayesian estimator for the shortest confidence interval at www.causascientia.org/math_stat/ProportionCI.html.

40. Whether agents or surrogates are grown in legitimate research labs only a small number of batches of limited size are produced. Similarly, except for state funded military programs, the illicit production of biological agents is expected to involve very limited quantities of material.

41. It is interesting to note that there is often a strong association between a particular laboratory and particular growth media, consistent with the observation that researchers

generally "stick to what works" when it comes to routine procedures such as culturing stocks of bacteria. It should also be understood that the large number and high degree of variety of media found in this survey are probably characteristic of the non-fastidious nature of *Bacillus* spp. Other organisms, e.g. *F. tularensis*, are known to have a much narrower range of culture media.

42. Kester, A.D.M, and Buntinx, F., "Meta-Analysis of ROC curves", Medical Decision Making, **20**, pp.430-439, (2000).

43. I.T. Jolliffe, *Principal Component Analysis*, (Springer-Verlag, New York, 1986).

Appendix 1. Source Attribution

Consider the case where a questioned sample is compared to an agent sample (the "source sample") found at a possible source lab, and a certain value of the test statistic Δ is obtained. Given the possibility of a false match, what is the likelihood that the questioned sample came from the suspect lab? Let L_0 be the hypothesis that the questioned sample was made in the same lab as the source sample, and $\overline{L}_0$ the hypothesis that it was made elsewhere. The odds that the questioned sample comes from the suspect lab *given* that the elemental vectors of the questioned and source sample pair differ by Δ is given by:

$$O(L_0|\Delta) = [P(\Delta|L_0)/P(\Delta|\overline{L}_0)] \bullet O(L_0) \qquad \text{(A1.1)}$$

where $O(L_0)$ are the prior odds that the questioned sample came from the suspect lab, which depend on other evidence. (For example, $O(L_0)$ might be considered very low if the laboratory in question followed a strict biosurety program and no other evidence pointed to possible theft of the existing stock material[1]. Similarly, $O(L_0)$ might be considered very high if the lab in question were found in the domicile of a person suspected of the crime for other reasons.) The quantity in brackets in equation (A1) is the likelihood ratio that expresses the strength of the evidence that the value of Δ provides to support the assertion that the lab is the source of the questioned sample.

$$LR(\Delta) = P(\Delta\ |L_0)/P(\Delta|\overline{L}_0) \qquad \text{(A1.2)}$$

Strictly speaking, a sample-matching test for bioagents provides evidence for two samples originating from the same batch of material (B_0) rather than from the same laboratory. Thus it is necessary to formally relate the likelihood ratio (A2) to the probabilities of match and false match given B_0 or $\overline{B}_0$. An important consideration in relating Δ, B_0 and L_0 is that the laboratory in question may have made many batches of the bioagent, and not all of them may be available for testing. Using the chain rule for conditional probabilities we can write:

$$P(\Delta|L_0) = P(\Delta|B_0L_0) \bullet P(B_0|L_0) + P(\Delta|\overline{B}_0L_0) \bullet P(\overline{B}_0|L_0) \qquad \text{(A1.3)}$$

and

$$P(\Delta|\overline{L}_0) = P(\Delta|B_0\overline{L}_0) \bullet P(B_0|\overline{L}_0) + P(\Delta|\overline{B}_0\overline{L}_0) \bullet P(\overline{B}_0|\overline{L}_0) \qquad \text{(A1.4)}$$

The definitions of the various probability functions that are factors in equations (A3) and (A4) are given in table A1.1.

These expressions refer to 3 possible sub-populations of samples:

$\{B_0L_0\}$ the set of pairs drawn from the same batch, made in the same laboratory;

$\{\overline{B}_0L_0\}$ the set of pairs drawn from the different batches, made in the same laboratory;

and $\{\bar{B}_0\bar{L}_0\}$ the set of pairs drawn from different batches made in different laboratories.

The fourth set $\{B_0\bar{L}_0\}$ is an empty set, since two samples made in different laboratories clearly cannot come from the same batch. It is important to note, that samples drawn from different batches can come from batches made by identical or different processes. In set notation:

$$\{\bar{B}_0L_0\} = \{\bar{B}_0P_0L_0\} \cup \{\bar{B}_0\bar{P}_0L_0\} \text{ and } \{\bar{B}_0\bar{L}_0\} = \{\bar{B}_0P_0\bar{L}_0\} \cup \{\bar{B}_0\bar{P}_0\bar{L}_0\}.$$

Using the values given in the table, note that:

$$P(\Delta|L_0) = P(\Delta|B_0L_0)\bullet(1/M_B) + P(\Delta|\bar{B}_0L_0)\bullet(1 - 1/M_B) \qquad \text{(A1.5)}$$

and

$$P(\Delta|\bar{L}_0) = P(\Delta|\bar{B}_0\bar{L}_0) \qquad \text{(A1.6)}$$

Table A1.1 Quantities used in the analysis of attribution.

Quantity	Definition	*A priori* value
$P(\Delta\|B_0L_0)$	Probability that the two samples will have the observed Δ value if they are from the same batch of material from the suspect lab	Empirically determined by measurements on samples made in the same laboratory
$P(\Delta\|\bar{B}_0L_0)$	Probability that the two samples will have the observed Δ value if they are not from the same batch of material, but are from the suspect lab	Empirically determined by measurements on samples made in the same laboratory, but from different batches and processes.
$P(B_0\|L_0)$	Probability that the questioned sample was drawn from the same batch as the suspect source sample, given that they both come from the same lab	$1/M_B$ where M_B is the number of independent batches made by the laboratory in question
$P(\bar{B}_0\|L_0)$	Probability that the questioned sample was not drawn from the same batch as the suspect source sample, given that they both come from the same lab	$1 - 1/M_B$
$P(\Delta\|B_0\bar{L}_0)$	Probability that the two samples will have the observed Δ value if they are from the same batch of material and they come from different labs	Undefined (The same batch of material cannot originate in different labs.)
$P(\Delta\|\bar{B}_0\bar{L}_0)$	Probability that the two samples will have the observed Δ value if they are from different batches of material and they come from different labs	Empirically determined by measurements on samples made in different laboratories, from different batches and processes
$P(B_0\|\bar{L}_0)$	Probability that the questioned sample was drawn from the same batch as the suspect source sample, given that they came from different labs	0
$P(\bar{B}_0\|\bar{L}_0)$	Probability that the questioned sample was not drawn from the same batch as the suspect source sample, given that they came from different labs	1

If the suspect lab made only a single batch of the material in question ($M_B = 1$) then the probability of observing Δ given that the samples came from the same lab is clearly the same as observing Δ if the samples came from the same batch. In this case the likelihood ratio in equation (A2) simply becomes:

$$LR(\Delta) = P(\Delta|B_0L_0)/P(\Delta|\overline{B}_0\overline{L}_0). \qquad (A1.7)$$

However, if the lab made many independent batches of material (i.e. $M_B >> 1$), the likelihood ratio will primarily depend on the probability of observing Δ among *different* batches made in the same lab, and

$$LR(\Delta) \approx P(\Delta|\overline{B}_0L_0)/P(\Delta|\overline{B}_0\overline{L}_0). \qquad (A1.8)$$

Because batches made in one lab are liable to be replicate batches while batches made in different laboratories are liable to be non-replicate even when made by the sample process, we expect

$$P(\Delta|\overline{B}_0L_0) > P(\Delta|\overline{B}_0\overline{L}_0) \qquad (A1.9)$$

for some range of Δ values. Thus it is still possible to have probative weight to a given value of Δ even if a lab made many batches of agent and some are not available for testing.

In summary, the probative value of a given value of Δ is bounded by two limits (A1.7) and (A1.8) representing the cases where we are certain the lab produced only one batch of the material in question, and where the lab may have produced many batches of material.

The ROC curves described in section 2 of the main text are defined in terms of the marginal probabilities

$$P(\Delta \leq \Delta_b|B_0L_0) = \int_0^{\Delta_b} P(\Delta'|B_0L_0)d\Delta', \qquad (A1.10)$$

$$P(\Delta \leq \Delta_b|\overline{B}_0L_0) = \int_0^{\Delta_b} P(\Delta'|\overline{B}_0L_0)d\Delta', \qquad (A1.11)$$

and

$$P(\Delta \leq \Delta_b|\overline{B}_0\overline{L}_0) = \int_0^{\Delta_b} P(\Delta'|\overline{B}_0\overline{L}_0)d\Delta'. \qquad (A1.12)$$

The likelihood ratios in (A1.7) and (A1.8) can therefore be expressed as the derivative of the relevant ROC curve, e.g.

$$P(\Delta|B_0L_0)/P(\Delta|\overline{B}_0L_0)= dP(\Delta \leq \Delta_b|B_0L_0)/dP(\Delta \leq \Delta_b|\overline{B}_0L_0) \qquad (A1.13)$$

The analysis above implies that there are 2 separate ROC curves that need to be determined in order to estimate the likelihood ratios (A7) and (A8). One of the ROC curves, $P(\Delta \leq \Delta_b|B_0L_0)$ *versus* $P(\Delta \leq \Delta_b|\overline{B}_0L_0)$, involve batches made in the same laboratory. A conservative point of view is that batches of material made at the same laboratory are more likely to be replicate batches, and to use only one process, so that these quantities are conservatively estimated from experiments on sets of replicate batches. The most conservative estimate of the marginal probability, $P(\Delta \leq \Delta_b|\overline{B}_0L_0)$, is obtained by having all replicate cultures made in one laboratory by one person, using the same lot of medium components or a pre-manufactured batch of growth medium, and carefully implemented quality control steps.

The other required ROC curve is $P(\Delta \leq \Delta_b|\overline{B}_0L_0)$ versus $P(\Delta \leq \Delta_b|\overline{B}_0\overline{L}_0)$, which deals with samples made in different laboratories. The latter quantity is conservatively estimated from measurements on non-replicate batches of material made independently, but by the same process. Note that the joint hypothesis B_0L_0 is equivalent to B_0 alone since the same batch of material must be made in the same laboratory. However, as noted above there is a distinction between multiple batches made at the same laboratory and batches made at different laboratories, i.e. between $\overline{B}_0L_0$ and $\overline{B}_0\overline{L}_0$. In section 4 of this report we follow the prescription suggested in this appendix that $\overline{B}_0L_0$ is conservatively represented by replicate batches, while $\overline{B}_0\overline{L}_0$ is conservatively represented by non-replicate batches, whether or not they are actually made in different laboratories.

In general, the finding that two samples were made by the same process (P_0) has far less probative value in associating the questioned sample with a given laboratory, unless the process can be uniquely associated with that laboratory. Expanding the likelihood ratio for a match in terms of P_0 rather than B_0, the analogues to equations (A1.3) and (A1.4) become:

$$P(\Delta|L_0) = P(\Delta|P_0L_0) \bullet P(P_0|L_0) + P(\Delta|\overline{P}_0L_0) \bullet P(\overline{P}_0|L_0) \qquad (A1.14)$$

and

$$P(\Delta|\overline{L}_0) = P(\Delta|P_0\overline{L}_0) \bullet P(P_0|\overline{L}_0) + P(\Delta|\overline{P}_0\overline{L}_0) \bullet P(\overline{P}_0|\overline{L}_0) \qquad (A1.15)$$

$P(P_0|L_0)$ is the probability that two samples were made by the same process if they were made in the same laboratory. If other evidence made it highly likely that the suspect lab used only a single process to make bioagent material (i.e. $P(P_0|L_0) \approx 1$) then the probability of observing Δ given that the samples came from the same lab is clearly the same as observing Δ if the samples were made by the same process (i.e. $P(\Delta|L_0) \approx P(\Delta|P_0L_0)$. Similarly, if it is highly unlikely that a different laboratory might use the same process to generate agent, then $P(P_0|\overline{L}_0) \approx 0$ and $P(\Delta|\overline{L}_0) \approx P(\Delta|\overline{P}_0\overline{L}_0)$. In this case the likelihood ratio in equation (A2) simply becomes:

$$LR(\Delta) = P(\Delta|P_0L_0)/P(\Delta|\overline{P}_0\overline{L}_0). \tag{A1.16}$$

Because the probability of obtaining a value of Δ with two different processes used by one lab is the same as the probability for processes used by *any* lab, we expect

$$P(\Delta|\overline{P}_0L_0) \approx P(\Delta|\overline{P}_0\overline{L}_0), \tag{A1.17}$$

and the likelihood ratio becomes, simply:

$$LR(\Delta) = P(\Delta|P_0)/P(\Delta|\overline{P}_0). \tag{A1.18}$$

In the more general case where it is possible that different processes may have been carried out in the same laboratory, the likelihood ratio is more difficult to calculate exactly, but it may be possible to argue that a process matching test still has probative value (i.e. $LR(\Delta) > 1$). This will be true if the $P(\Delta|\overline{P}_0L_0)$ and $P(\Delta|\overline{P}_0\overline{L}_0)$ are sufficiently small compared to $P(\Delta|P_0L_0)$ and $P(\Delta|P_0\overline{L}_0)$ and we accept the proposition that $P(P_0|L_0) > P(P_0|\overline{L}_0)$.

Table A1.2. Probabilities associated with process matching.

Probability	Definition
$P(\Delta|P_0L_0)$	Probability that the two samples will have the observed Δ value given that the samples were made by the same process in the same laboratory
$P(\Delta|P_0\overline{L}_0)$	Probability that the two samples will have the observed Δ value given that the samples were made by the same process in different laboratories
$P(\Delta|\overline{P}_0L_0)$	Probability that the two samples will have the observed Δ value given that the samples were made by *different* processes in the same laboratory
$P(\Delta|\overline{P}_0\overline{L}_0)$	Probability that the two samples will have the observed Δ value given that the samples were made by *different* processes in *different* laboratories
$P(P_0|L_0)$	Probability that two samples were made by the same process, given that they were made in the same laboratory
$P(\overline{P}_0|L_0)$	Probability that two samples were made by *different* processes, given that they were made in the same laboratory
$P(P_0|\overline{L}_0)$	Probability that two samples were made by the same process, given that they were made in *different* laboratories
$P(\overline{P}_0|\overline{L}_0)$	Probability that two samples were made by *different* processes, given that they were made in *different* laboratories

Another case where the likelihood ratio takes a simpler form is where there is essentially only one process that can be used to make the bioagent. Such a situation may exist for certain fastidious bacterial agents. In this case,

$$P(P_0|L_0) \approx P(P_0|\overline{L}_0) \approx 1 \qquad \text{(A1.19)}$$

And the likelihood ratio reduces to :

$$LR(\Delta) = P(\Delta|P_0L_0)/P(\Delta|P_0\overline{L}_0). \qquad \text{(A1.20)}$$

Since we expect that the probability of finding a match between two samples made by a given process in the same lab is higher than the probability of a match between samples made in different labs, even when made by the same process, the likelihood ratio will be greater than 1 and a process match will have some probative value towards attribution.

To summarize this discussion, the relevant conditional probabilities associated with process matching are given in Table A1.2.

[1] There are several reasons why a legitimate bio-defense laboratory might have small amounts of biological agent at hand. First, the material might be required to challenge vaccine candidates or to test the efficacy of other kinds of treatments for exposure to the agent. Second, the material might be required for "gold standard" testing of new detection technologies. Finally, the material may have been generated as part of a threat assessment program and has been retained for reference purposes. Although recent increases in biosurety measures have made it less probable that such material could be stolen for purposes of bioterrrorism, it is nonetheless a possibility that cannot be ignored in practice.

Appendix 2. Subpopulation size in the 3x3 partial factorial design

In section 4 the sample set is generated by drawing 3 replicate samples from each of the 18 separate batches, leading to 54 individual samples for analysis. This leads to 1431 possible pairwise comparisons of the elemental vectors. (If there are n samples then the number of pairwise comparisons is 1/2n(n-1).) This set of pairwise sample comparisons can be broken down into sets of pairs drawn from the various categories described in appendix 1: e.g. pairs drawn from the same batch (B_0L_0), from different batches produced in different labs using the same process ($\overline{L}_0P_0$), etc.

First consider pairwise comparisons of samples produced in the same laboratory. Table A2.1 represents the matrix of pairwise comparisons of samples produced by 3 different processes at a given laboratory.

Table A2.1. Matrix of pairwise comparisons for samples produced in the same laboratory.

		Process 1			Process 2		Process 3
		Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
Process 1	Batch 1	L_0P_0	L_0P_0	L_0P_0	$L_0\overline{P}_0$	$L_0\overline{P}_0$	$L_0\overline{P}_0$
	Batch 2		L_0P_0	L_0P_0	$L_0\overline{P}_0$	$L_0\overline{P}_0$	$L_0\overline{P}_0$
	Batch 3			L_0P_0	$L_0\overline{P}_0$	$L_0\overline{P}_0$	$L_0\overline{P}_0$
Process 2	Batch 4				L_0P_0	L_0P_0	$L_0\overline{P}_0$
	Batch 5					L_0P_0	$L_0\overline{P}_0$
Process 3	Batch 6						L_0P_0

In the cells along the diagonal (shaded pink) we are comparing samples from the same batch. Thus, there are three unique sample pairs per cell. The total number of such comparison pairs in the entire set is given by 3 pairs per cell x 6 cells per laboratory x 3 laboratories = 54 pairs.

The blue shaded cells represent pairs drawn from different batches made by the same process. There are 9 unique pairings per cell, leading to 9 x 4 x 3 = 108 such sample pairs in the entire set. Thus the total number of pairs drawn from the same process in the same laboratory (L_0P_0) is 108 + 54 = 162.

The remaining 11 cells (yellow) represent pairs drawn from batches made by different processes, for which there are 9 x 11 x 3 = 297 pairs.

For each process, we can consider the comparison of samples made by the same process in different laboratories. These are given in Table A2.2.

Each of the 11 cells represent 9 unique pairwise comparisons, so for the 3 processes we have 9 x 11 x 3 = 297 pairs in the set $\{\overline{L}_0P_0\}$.

Table A2.2. Pairwise comparisons of samples made in different laboratories using the same process.

		Lab1			Lab 2		Lab 3
		Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
Lab 1	Batch 1				$\overline{L}_0P_0$	$\overline{L}_0P_0$	$\overline{L}_0P_0$
	Batch 2				$\overline{L}_0P$	$\overline{L}_0P$	$\overline{L}_0P$
	Batch 3				$\overline{L}_0P$	$\overline{L}_0P$	$\overline{L}_0P$
Lab 2	Batch 4						$\overline{L}_0P$
	Batch 5						$\overline{L}_0P$
Lab 3	Batch 6						

For each pair of processes, we can construct the matrix describing the pairs of samples comparing different processes at different laboratories. This is shown in Table A2.3 for processes 1 and 3. There are 25 cells in this matrix, each representing 9 unique pairwise comparisons. Since there are 3 unique pairs of processes, the total number of comparisons in the set $\{\overline{P}_0\overline{L}_0\}$ is 25 x 9 x 3 = 675.

Table A2.3. Pairwise comparisons between processes 1 and 3 in different laboratories.

			Process 1					
			Lab 1			Lab 2		Lab3
			Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
Process 3	Lab1	Batch 1				$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$
		Batch 2				$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$
	Lab 2	Batch 3	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$			$\overline{P}_0\overline{L}_0$
	Lab 3	Batch 4	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	
		Batch 5	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	
		Batch 6	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	$\overline{P}_0\overline{L}_0$	

Table A2.4 summarizes the numbers of members in the various sub-populations for the process-laboratory comparisons.

Table A2.4

	P_0	$\overline{P}_0$	Total
L_0	162	297	459
$\overline{L}_0$	297	675	972
Total	459	972	1431

Thus, for example, the probability of drawing at random a pair of samples representing the same process carried out at the same laboratory from this set is 162/1431 or 11%.

For each process, we can summarize the various comparisons among batches as in Table A2.5. The structure of this matrix is the same as that of A2.1. The diagonal cells, represent comparisons between samples drawn from the same batch. Each diagonal cell contains 3 pairwise comparisons among the 3 replicate samples drawn from that batch. Thus, the set $\{B_0L_0\}$ (≡ $\{B_0P_0L_0\}$) contains 3 x 6 x 3 = 54 pairs. Similarly, the blue shaded cells represent comparisons among samples from different batches made by the same process at the same lab $\{\overline{B}_0P_0L_0\}$. By analogy with table A2.1 there are 108 pairs in this set. Finally the yellow shaded cells in Table A2.5 represent a total of 297 pairwise comparisons between samples drawn from different batches made by the same process at different laboratories $\{\overline{L}_0P_0\overline{B}_0\}$.

Table A2.5. Pairwise comparisons between samples drawn from batches made by the same process.

		Lab 1			Lab 2		Lab 3
		Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
Lab 1	Batch 1	$L_0P_0B_0$	$L_0P_0\overline{B}_0$	$L_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$
	Batch 2		$L_0P_0B_0$	$L_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$
	Batch 3			$L_0P_0B_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$
Lab 2	Batch 4				$L_0P_0B_0$	$L_0P_0\overline{B}_0$	$\overline{L}_0P_0\overline{B}_0$
	Batch 5					$L_0P_0B_0$	$\overline{L}_0P_0\overline{B}_0$
Lab 3	Batch 6						$L_0P_0B_0$

The numbers of members in each of the sets defined by Table A.2.5 is summarized in Table A.2.6.

Table A2.6

	P_0B_0	$P_0\overline{B}_0$	Total
L_0	54	108	162
$\overline{L}_0$	0	297	297
Total	54	405	459

Note that the following relationships hold:

$$\{L_0B_0\} \equiv \{B_0P_0L_0\} \tag{A2.1}$$

$$\{L_0\overline{B}_0\} \equiv \{\overline{B}_0P_0L_0\} \cup \{\overline{B}_0\overline{P}_0L_0\};\ \{\overline{L}_0B_0\} = \varnothing \tag{A2.2}$$

$$\{\overline{L}_0\overline{B}_0\} \equiv \{\overline{B}_0P_0\overline{L}_0\} \cup \{\overline{B}_0\overline{P}_0\overline{L}_0\} \tag{A2.3}$$

and

$$\{L_0P_0\} \equiv \{B_0P_0L_0\} \cup \{\overline{B}_0P_0L_0\} \tag{A2.4}$$

$$\{L_0\overline{P}_0\} \equiv \{B_0\overline{P}_0L_0\} \cup \{\overline{B}_0\overline{P}_0L_0\};\ \text{but}\ \{B_0\overline{P}_0L_0\} = \varnothing \tag{A2.5}$$

$$\{\overline{L}_0P_0\} \equiv \{B_0P_0\overline{L}_0\} \cup \{\overline{B}_0P_0\overline{L}_0\};\ \text{but}\ \{B_0P_0\overline{L}_0\} = \varnothing \tag{A2.6}$$

$$\{\overline{L}_0\overline{P}_0\} \equiv \{B_0\overline{P}_0\overline{L}_0\} \cup \{\overline{B}_0\overline{P}_0\overline{L}_0\};\ \text{but}\ \{B_0\overline{P}_0\overline{L}_0\} = \varnothing. \tag{A2.7}$$

Therefore,

$$\{L_0\bar{B}_0\} \equiv \{\bar{B}_0P_0L_0\} \cup \{\bar{P}_0L_0\} \tag{A2.8}$$

$$\{\bar{L}_0\bar{B}_0\} \equiv \{P_0\bar{L}_0\} \cup \{\bar{P}_0\bar{L}_0\} = \{\bar{L}_0\}. \tag{A2.9}$$

Clearly, the probability of randomly drawing two samples from the same batch from the entire set of samples is 54/1431 or 3.8%. The probability of drawing a pair of samples that were made in the same lab, but *not* in the same batch is given by (108 + 297)/1431 and the probability of drawing a pair of samples that were grown in different laboratories is 972/1431. When determining the marginal probabilities $P(\Delta \leq \Delta_b|\bar{B}_0L_0)$ and $P(\Delta \leq \Delta_b|\bar{B}_0\bar{L}_0)$ in order to generate ROC curves, it is important to remember that batches from different processes are included in the sample sets $\{\bar{B}_0L_0\}$ and $\{\bar{B}_0\bar{L}_0\}$, as indicated in (A2.8) and (A2.9).