United States Government Accountability Office

Report to Congressional Requesters

December 2014

ANTHRAX

Agency Approaches to Validation and Statistical Analyses Could Be Improved

GAO Highlights

Highlights of GAO-15-80, a report to congressional requesters

Why GAO Did This Study

In 2001, the FBI investigated an intentional release of *B. anthracis*, a bacterium that causes anthrax, which was identified as the Ames strain. Subsequently, FBI contractors developed and validated several genetic tests to analyze *B. anthracis* samples for the presence of certain genetic mutations. The FBI had previously collected and maintained these samples in a repository.

GAO was asked to review the FBI's genetic test development process and statistical analyses. This report addresses (1) the extent to which these genetic tests were scientifically verified and validated; (2) the characteristics of an adequate statistical approach for analyzing samples, whether the approach used was adequate, and how it could be improved for future efforts; and (3) whether any remaining scientific concerns regarding the validation of genetic tests and statistical approaches need to be addressed for future analyses. GAO reviewed agency and contractor documentation, conducted literature reviews, and conducted statistical analyses of the repository data. GAO's review focused solely on two aspects of the FBI's scientific evidence: the validation of the genetic tests and the statistical approach for the analyses of the results. GAO did not review and is not taking a position on the conclusions the FBI reached when it closed its investigation in 2010.

What GAO Recommends

GAO recommends that the FBI develop a framework for validation and statistical approaches for future investigations. The FBI agreed with our recommendations.

View GAO-15-80. For more information, contact Timothy M. Persons, Chief Scientist, at (202) 512-6412 or personst@gao.gov.

December 2014

ANTHRAX

Agency Approaches to Validation and Statistical Analyses Could Be Improved

What GAO Found

After the 2001 Anthrax attacks, the genetic tests that were conducted by the Federal Bureau of Investigation's (FBI) four contractors were generally scientifically verified and validated, and met the FBI's criteria. However, GAO found that the FBI lacked a comprehensive approach—or framework—that could have ensured standardization of the testing process. As a result, each of the contractors developed their tests differently, and one contractor did not conduct verification testing, a key step in determining whether a test will meet a user's requirements, such as for sensitivity or accuracy. Also, GAO found that the contractors did not develop the level of statistical confidence for interpreting the testing results for the validation tests they performed. Responses to future incidents could be improved by using a standardized framework for achieving minimum performance standards during verification and validation, and by incorporating statistical analyses when interpreting validation testing results.

GAO identified six characteristics of a statistical framework that can be applied for analyzing scientific evidence. When GAO compared the approach the FBI used to this framework, it found that that the FBI's approach could have been improved in three of six areas. First, the FBI's research did not provide a full understanding of the methods and conditions that give rise to genetic mutations used to differentiate between samples of *B. anthracis*. Second, the FBI did not institute rigorous controls over the sampling procedures it used to build the repository of *B. anthracis* samples. Third, the FBI did not include measures of uncertainty to strengthen the interpretation of the scientific evidence. GAO found that since 2001 the FBI has taken some steps to build formal forensic statistical expertise. The FBI's approach to future incidents could benefit from including such expertise early in an investigation.

The lack of an understanding of how bacteria change (mutate) in their natural environment and in a laboratory is a key scientific gap that remains and could affect testing conducted in future incidents. Specifically, the significance of using such mutations as genetic markers for analyzing evidentiary samples to determine their origins is not clear. This gap affects both the development of genetic tests targeting such mutations and statistical analyses of the results of their use on evidentiary samples. The Department of Homeland Security is currently funding some research on genetic changes in bacteria and genome sequencing methods, among others. Such research is a step in the right direction since the FBI is planning to use genome sequencing methods in future investigations. However, because this research may not be complete for several more years, the extent to which it will close this gap is not known.

Contents

Tables

Figures

Abbreviations

APHIS	Animal and Plant Health Inspection Service
ASA	American Statistical Association
B. anthracis	*Bacillus anthracis*
CBI	Commonwealth Biotechnologies
CBSU	Chemical and Biological Service Unit
CDC	Centers for Disease Control and Prevention
DHS	U.S. Department of Homeland Security
DNA	deoxyribonucleic acid
DOD	U.S. Department of Defense
DOE	U.S. Department of Energy
FBI	Federal Bureau of Investigation
FBIR	FBI Repository of Ames Samples
FDA	Food and Drug Administration
IITRI	Illinois Institute of Technology Research Institute
INDEL	Insertion/deletion
JGI	Joint Genome Institute
LLNL	Lawrence Livermore National Laboratory
LOD	limit of detection
MRI	Midwest Research Institute
NAS	National Academy of Sciences
NAU	Northern Arizona University
NBACC	National Biodefense Analysis and Countermeasures Center
NBFAC	National Bioforensics Analysis Center
NIST	National Institute of Standards and Technology
NRC	National Research Council
NSTC	National Science and Technology Council
OSTP	Office of Science and Technology Policy
PCR	polymerase chain reaction
SOP	standard operating procedure
SWGDAM	Scientific Working Group on DNA Analysis Methods
SWGMGF	Scientific Working Group on Microbial Genetics and Forensics
TIGR	The Institute for Genomic Research
USAMRIID	United States Army Medical Research Institute for Infectious Diseases

GAO U.S. GOVERNMENT ACCOUNTABILITY OFFICE

441 G St. N.W.
Washington, DC 20548

December 19, 2014

The Honorable Yvette Clark
Ranking Member
Subcommittee on Cybersecurity, Infrastructure
Protection, and Security Technologies
Committee on Homeland Security
House of Representatives

The Honorable Rush Holt
House of Representatives

The Honorable Jerrold Nadler
House of Representatives

In September and October 2001, letters laced with *Bacillus anthracis* Ames strain (*B. anthracis*) were mailed through the U.S. postal system to two U.S. senators and members of the media. Public health entities, county and state public health departments and the Centers for Disease Control (CDC), in collaboration with the Federal Bureau of Investigation (FBI) and the U.S. Postal Inspection Service, took the initial lead in the investigation.[1] The FBI worked in collaboration with public health entities and U.S. Postal Inspection Service when a deliberate act was eventually suspected, following the identification of the attack letters and the first victim in Florida. In 2007, the FBI determined that the spores in the letters were derived from a single spore-batch of the Ames strain in a flask called “RMR-1029.”

In 2008, the FBI asked the National Research Council (NRC) of the National Academy of Sciences (NAS) to review the scientific approaches it had used to support its conclusions. During the NRC committee’s deliberation, the FBI announced on February 19, 2010, that it was closing the case, having concluded that a scientist at the United States Army Medical Research Institute for Infectious Diseases (USAMRIID) had perpetrated the attack alone.[2] In February 2011, the NAS issued its

[1] The Environmental Protection Agency was later involved.

[2] U.S. Department of Justice, *Amerithrax Investigative Summary*, Washington, D.C., February 19, 2010.

report, concluding that "it is not possible to reach a definitive conclusion about the origins of the *B. anthracis* in the mailing based on the available scientific evidence alone."[3] The NAS report detailed many methodological and organizational problems in the scientific portion of the FBI's investigation, known by the case name *Amerithrax*.

Because several scientific and technical issues were not covered in the scope of the NAS study, you asked us to conduct an independent technical evaluation of the scientific approaches used in support of the FBI's investigation, focusing on certain issues. We reviewed the NAS findings and conclusions and determined that, since NAS did not report in depth how the genetic assays (or tests) used to screen the repository were scientifically verified and validated, additional evaluation of the requirements and procedures used for doing so could be informative in developing scientific methods for future investigations. The NAS report stated that, "Although the committee lauds and supports the effort dedicated to the development of well-validated assays and procedures, looking toward the future, these processes need to be more efficient." Further, the NAS report included several findings related to the statistical analyses of the repository data and identified challenges concerning the identification and collection of the repository samples. Therefore, a review of the comprehensive statistical approach could help further clarify and expand on the impact of the NAS conclusions and provide useful insight for applying statistical approaches to future investigations. Consequently, this report addresses the following three questions:

1. To what extent were the genetic assays used to screen the FBI repository of Ames samples scientifically verified and validated?[4]
2. What are the characteristics of an adequate statistical approach for analyzing the repository samples and to what extent was the statistical approach used adequate? If not adequate, how could this approach be improved for future efforts?
3. What remaining scientific concerns and uncertainties, if any, regarding the validation of genetic tests and statistical approaches will need to be addressed in future analyses? What additional research, if any,

[3] National Research Council, *Review of the Scientific Approaches Used during the FBI's Investigation of the 2001 Anthrax Letters* (Washington, D.C.: Feb. 2011).

[4] Assay and test are synonymous. In this report we generally refer to the genetic assays as genetic tests.

would be helpful in resolving such scientific uncertainties in any future investigation?

The scope of our work was limited to a review of the scientific methods employed to verify and validate the genetic tests used to screen the FBI's repository of Ames *B. anthracis* samples, the procedures used to identify and collect samples of Ames *B. anthracis* in the creation of the FBI's repository, and the statistical analyses and interpretation of the results of the genetic tests. Thus, we did not address any other scientific methods or any of the traditional investigative techniques used to support the FBI's conclusions in this case, and we take no position on the conclusions the FBI reached when it closed its investigation in 2010.

To meet our objectives, we reviewed pertinent agency and FBI contractor documentation on the verification and validation of the genetic tests used to screen the FBI's repository of Ames samples. We also reviewed scientific literature and agency and industry guidelines to determine the essential phases in approaches to developing genetic tests. We reviewed existing FBI standards and guidelines for verifying and validating microbial forensics methods. To determine the extent to which the genetic tests were verified and validated, we reviewed the results of the contractors' validation tests, the FBI's requirements for the validation, and the contractors' approaches for developing their genetic tests against an approach, or framework, for validation, encompassing the essential phases that we identified. We interviewed officials and scientists at the FBI, its contractors, and others regarding the development and validation of the genetic tests.

To determine the adequacy of the FBI's statistical approach, we conducted a literature review to identify the characteristics of an adequate statistical approach, analyzed agency and FBI contractor documentation, and interviewed FBI and laboratory officials. We conducted an analysis of the repository data to examine the effect of data trimming assumptions and additional estimates of false negative rates on the conclusions of the FBI's statistical analyses.[5]

[5] To illustrate the potential effect that uncertainty could have had on the interpretation of the results, we conducted an analysis using the estimates of false negative rates obtained from the additional replicate testing, combined with a sensitivity analysis accounting for the decision to restrict the statistical analyses to the 947 samples that contained no inconclusive or variant results. Appendix I has more details.

To identify scientific gaps and associated research related to questions 1 and 2, we reviewed agency and contractor documents and related scientific literature/reports and research agendas. We interviewed officials, statisticians, and scientists from the Department of Homeland Security (DHS), the FBI, and others regarding scientific gaps. We met with entities that were involved in the validation of the genetic tests and research addressing scientific gaps, such as the FBI and DHS contractors, and others involved in the repository testing. Finally, we used technical and scientific guidance we received from experts on the technologies and statistical approaches used in the FBI's investigation, and we read their comments on our draft report. We selected these experts for their expertise in public health and microbial forensics. Further details of our scope and methodology are in appendix I.

We conducted this performance audit from January 2013 to November 2014 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Background

The Attack History

In September and October 2001, at least seven envelopes containing significant quantities of *B. anthracis* spores were mailed through the U.S. postal system to two senators at their congressional offices in the District of Columbia and to media organizations in New York City and Boca Raton, Florida. According to the FBI, the evidence supports the conclusion that the mail attacks occurred on two separate occasions. The two letters of the first attack were postmarked September 18, 2001, and sent to NBC News and the New York Post, both in New York City. Three weeks later, two letters postmarked October 9, 2001, were mailed to two senators—Thomas Daschle and Patrick Leahy—at their Washington, D.C., offices. Other letters were sent to ABC, CBS, and American Media, Inc. Hard evidence of the attacks surfaced on October 3, 2001, when Robert Stevens, an American Media Inc. employee who worked in Boca Raton, Florida, was diagnosed as having contracted inhalational anthrax, from which he later died. However, because a contaminated envelope or package was not recovered in Florida, the agencies could not initially establish how the *B. anthracis* spores were delivered. According to the

Postal Service, the combination of the Florida incident and the opening of the letter to Senator Tom Daschle on October 15 established the link to the U.S. mail system. At least 22 victims contracted anthrax as a result of the mailings. Eleven individuals developed inhalational anthrax, and another 11 developed cutaneous infections. Five of the inhalational anthrax victims died from their infections.

The attack highlighted the need for enhanced capabilities for full forensic exploitation and interpretation of microbial evidence from acts of bioterrorism.[6] Ideally, forensic evidence obtained in an investigation is sufficient to support conclusions about the culpability of a group, an individual, or the source of material used in such an act.[7] Forensic evidence is used to support conclusions by classifying evidence into one of several categories that distinguish possible sources from one another. While classification does not unequivocally demonstrate a connection with an individual or a single source, it can be used to reduce the number of possible sources and thus can provide important leads in an investigation.

The development and application of microbial forensics was essential to the FBI's scientific investigation, which relied heavily on genetics and comparative genomics to classify the spore materials used in the attack, reduce the number of possible sources and suspects, and provide investigative leads. In fact, according to the NAS, this investigation accelerated the development of the then nascent field of microbial forensics.[8]

The FBI's Genetic Analyses of the Attack Spores

The FBI's investigation, assisted by government, university, and commercial laboratories, was an effort to develop the physical, chemical, genetic, and forensic profiles of the anthrax spores in the letters and envelopes used in the attacks. The investigation employed myriad

[6] Microbial forensics characterizes, analyzes, and interprets microbial evidence for attribution purposes. The field has grown from the multidisciplinary fields of genomics, microbiology, and forensics, among others.

[7] Referred to as "individualization," and sometimes as "matching," the properties of evidence to a particular individual or source.

[8] Microbial forensics has also been referred to as "bioforensics" and "forensic microbiology."

traditional and novel investigative and scientific methods. The scientific methods involved efforts to develop the physical, chemical, genetic, and forensic profiles of the anthrax spores and letters and envelopes used in the attacks so as to identify the source of the spores. The FBI faced many difficult and complex scientific challenges over the course of this investigation, according to the NAS. New microbial forensic methods were developed and implemented over several years, and some of them provided valuable evidence and significant leads in the case. For example, according to the FBI, new methods to determine the source of the growth media for the mailed spores were inconclusive while the use of the genetic mutations provided an investigative lead.

By October 2001, CDC had identified the microorganism used in the attack as *Bacillus anthracis (B. anthracis)*. This was a key step in the classification of the microorganism used in the attack letters and was one of the first scientific findings that allowed the FBI to begin to reduce the number of possible sources of the spores. *B. anthracis* is a gram positive, rod-shaped bacterium that causes the disease anthrax. It is a member of the larger genus *Bacillus* that includes other commonly found species, such as *B. cereus*, *B. subtilis*, and *B. thuringiensis*. *B. anthracis*, a species of *Bacillus* that can be found on all continents except Antarctica, typically shows little genetic variation among isolates. However, during the investigation scientific methods were being developed that allowed scientists to find some genetic differences among natural isolates of *B. anthracis*.[9] Applying these methods allowed the FBI to refine the classification of the spores used in the attack and further reduce the number of possible sources of the spores.

In October 2001, scientists working with the FBI identified the specific strain of *B. anthracis* used in the attack as the Ames strain.[10] Originally isolated from a dead cow in Texas in 1981, the Ames strain is uncommon in nature. It was shipped to USAMRIID and, over time, it was shared with laboratories around the world. The identification of the Ames strain significantly reduced the number of possible sources of the material used

[9] An isolate is a population of microbial cells in pure culture derived from a single colony on an isolation plate and identified to the species level.

[10] Even in the most homogeneous species, some differences are usual in genome sequences among populations. Although few in number, these differences are sufficient to characterize subgroups, or “strains.”

in the attack letters. In fact, this scientific evidence allowed the FBI to focus its investigation on the limited number of laboratories that had had access to the Ames strain before the attacks (see figure 1).

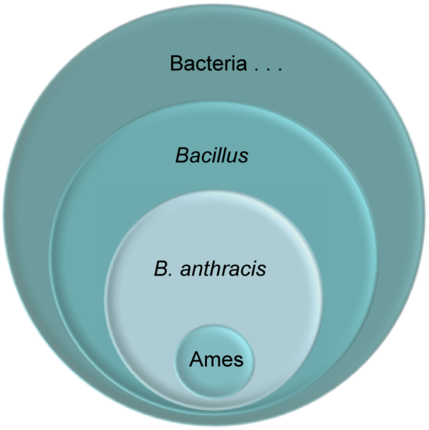
Figure 1: The Classification and Reduction of Possible Sources of Spores in the 2001 Anthrax Attack Letters

Bacteria: Found worldwide: numerous species.

Bacillus: Found on 6 of the 7 continents, in both nature and laboratories: several species.

B. anthracis: Found on 6 of the 7 continents, in both nature and laboratories.

Ames *B. anthracis:* Found rarely in nature. The FBI collected a repository of samples from more than 1,000 laboratory isolates.



Source: GAO. | GAO-15-80

Note: Diagram is not to scale.

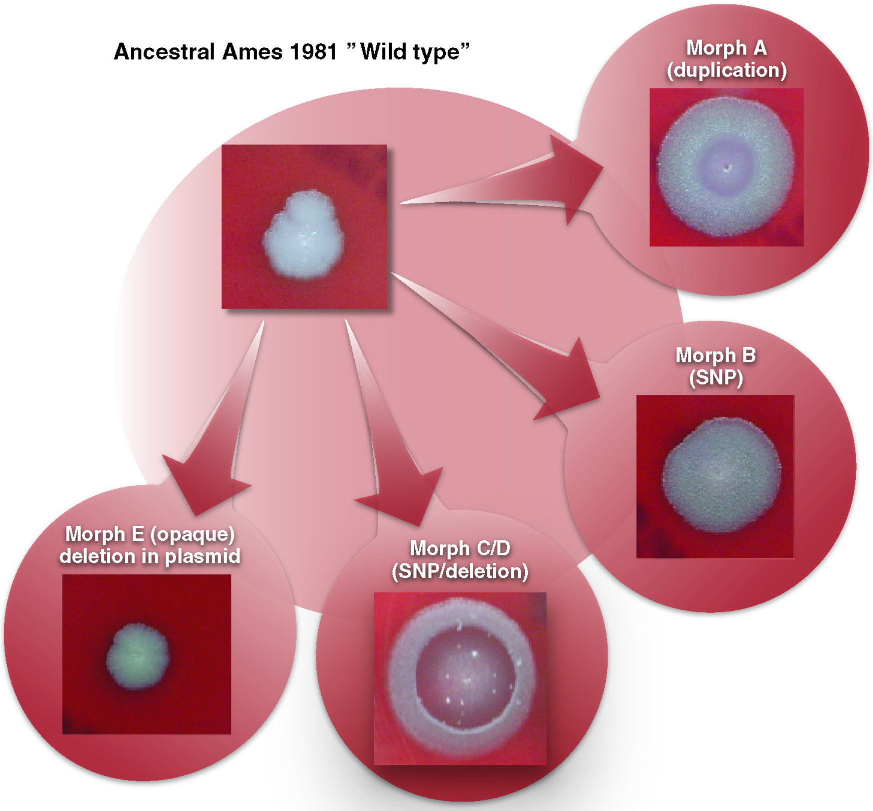
While classifying the spore material as the Ames strain was instrumental in reducing the number of possible sources, it was not sufficient by itself to definitively identify the source of the material used in the attack as a single laboratory, flask, or person. The FBI then sought to identify additional characteristics of the spores used in the attack that could further discriminate between possible sources of the Ames strain.

Scientists from the Department of Defense (DOD) tested samples of the spore materials found in the letters and identified several morphological variants (or morphs).[11] A laboratory technician who had grown (cultured) the spores from the letters over an extended period observed that a small percentage of the colonies differed in texture, color, and growth patterns

[11] Morphological variants are individual organisms that differ in observable physical or biochemical characteristics. These characteristics may be determined by environmental influences or a combination of the genetic makeup of the individual and environmental influences.

from those typical for the Ames strain of *B. anthracis*, referred to in figure 2 as the "wild type."[12]

Figure 2: Ancestral Ames Strain and Types of Morphs Found in the Evidence from the 2001 Anthrax Attack


Ancestral Ames 1981 "Wild type"
Morph A (duplication)
Morph B (SNP)
Morph E (opaque) deletion in plasmid
Morph C/D (SNP/deletion)

Sources: GAO and photographs courtesy of USAMRIID. I GAO-15-80

Note: SNPs and INDELs (insertions or deletions) are small differences between genomes. A duplication is a kind of insertion where a specific sequence of the deoxyribonucleic acid (DNA) has been repeated. These changes in genetic sequence are an important part of evolution.

[12] A bacterial colony is a visible cluster of microorganisms that originate from a single cell, thereby constituting clonal bacteria that are all alike genetically.

In an effort to identify the source of the letters, investigators and FBI scientists began to evaluate whether they could first identify and characterize these morphs genetically and then determine whether any of them were present in the repository of Ames samples. This involved genome sequencing to identify whether specific deoxyribonucleic acid (DNA) sequences underlay the morphs. Eventually, as shown in figure 2, the morphs were associated with several types of genetic mutations: duplications, single nucleotide polymorphisms (SNP), and deletions, referred to as INDELS.

Afterward, over several years, outside contractors' laboratories developed and validated several genetic tests to analyze the *B. anthracis* samples for the presence of certain genetic mutations.[13] Specifically, the testing revealed the presence or absence in a sample of a specific DNA sequence (that is, the genetic mutation) associated with a given morph.[14] The FBI contractors generally referred to the tests they developed as A1, A3, D, and E. Commonwealth Biotechnologies (CBI) developed the two A tests (A1 and A3), which targeted two different DNA sequences; and the Institute for Genomic Research (TIGR) developed the E test, targeting another DNA sequence. However, unlike the others, the Illinois Institute of Technology Research Institute (IITRI) and Midwest Research Institute (MRI) both developed a test targeting the same DNA sequence. For clarity, we refer to the IITRI-developed test as D-1 and the MRI-developed tests as D-2.

- **Genetic test A1:** detects the presence of a specific duplicated DNA sequence associated with morph A;
- **Genetic test A3:** detects a different duplicated DNA sequence from that targeted by A1 but also associated with morph A;
- **Genetic test D-1:** detects the presence of a specific deleted DNA sequence associated with morph D;

[13] The contractors did not develop genetic tests for all the genetic mutations associated with the morphs illustrated in figure 2. For example, the A morph was found to have three associated genetic mutations, duplications referred to as A1, A2, and A3. CBI was unable to develop a genetic test to detect the presence of the A2 mutation. Further, although three of the contractors attempted to develop a genetic test that would detect the SNP associated with the B morph— they were unsuccessful. Finally, two genetic mutations were associated with the E morph; genetic test (E) was developed to target one of them.

[14] DNA sequence is the specific order of the four nucleotides in a DNA molecule, sometimes referred to as base pairs because of the manner in which they form pairs.

- **Genetic test D-2:** detects the presence of the same deleted DNA sequence associated with morph D as that targeted by the D-1 test; and
- **Genetic test E:** detects a deleted DNA sequence associated with morph E.

In 2002, the FBI began collecting samples from laboratories in possession of the Ames strain to compare them with the material used in the attack. A grand jury issued subpoenas to 16 domestic laboratories and the FBI requested submissions from 3 foreign laboratories that investigators had determined possessed the Ames strain. The subpoenas required each laboratory to identify and submit two representative samples from each distinct stock of the Ames strain it held.[15] The subpoena included instructions to the laboratories on how to identify, select, and submit samples to the FBI. Laboratories were required to ship sample submissions to DOD scientists at USAMRIID for preparation and entry into the FBI repository of Ames samples.

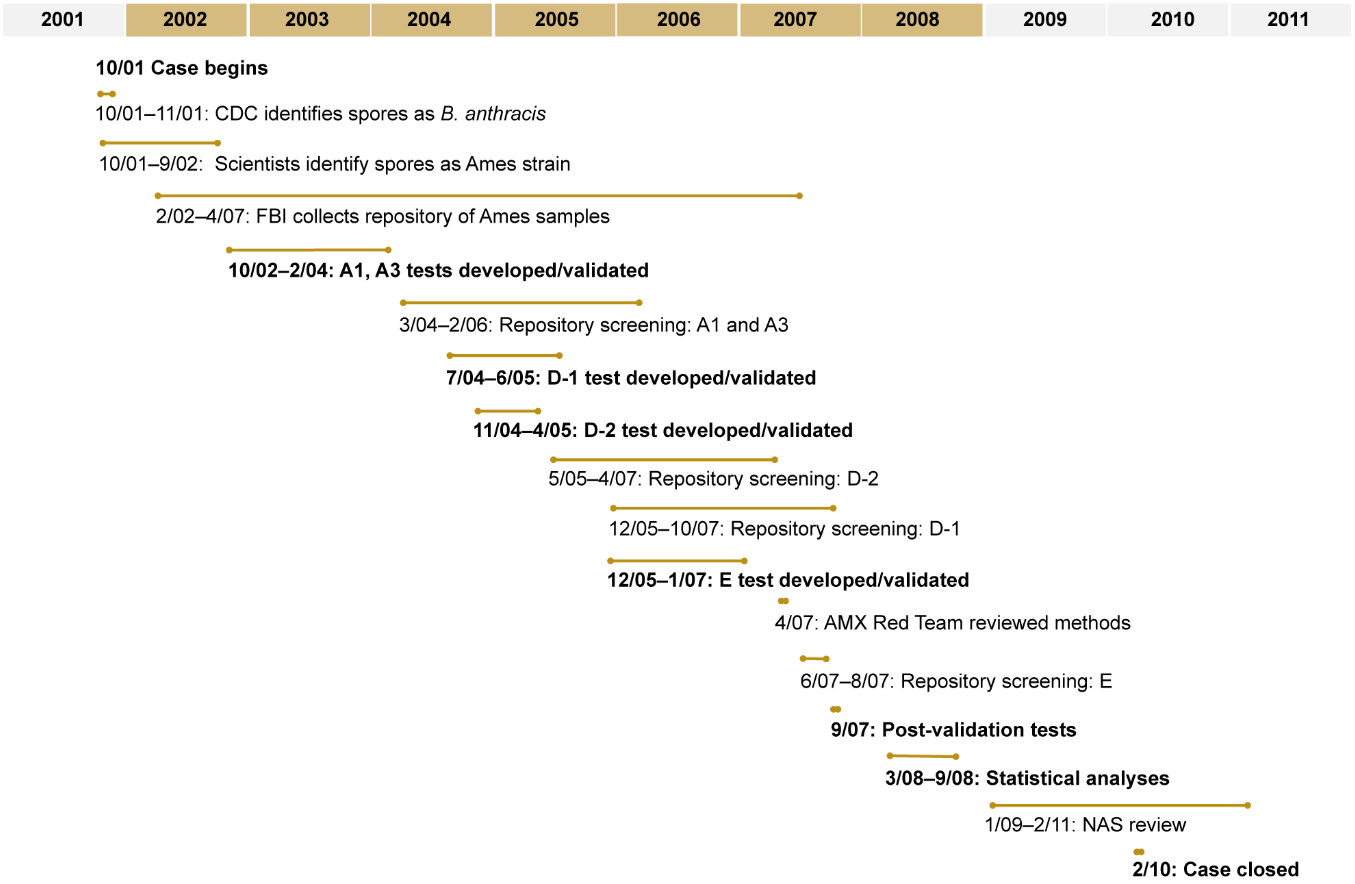
In addition to the samples submitted in response to the subpoena, searches were conducted at three domestic laboratories to ensure that samples were taken from each stock of Ames strain in those facilities. The FBI assembled a repository of 1,070 Ames strain samples, of which 1,059 were viable.[16] From 2004 through 2007, each of the 1,059 viable repository submissions was compared to the evidentiary material using the five genetic tests (see figure 3). The results of the genetic testing indicated that only 8 of the 1,059 FBI repository Ames samples tested positive for the presence of the four genetic mutations originally found in the anthrax letter evidence.[17]

[15] Stock refers to the microorganism maintained, for example, in a laboratory under conditions intended to keep it viable for subculture into fresh medium.

[16] Viability refers to, for example, the ability of cells to actively grow and form visible colonies on solid growth media.

[17] In this report, when we discuss the genetic mutations that the genetic tests were intended to identify, we are generally referring to the specific DNA sequences associated with the morphs. However, the FBI, NAS, and others often use “morph” and genetic “mutation” or genetic “variant” interchangeably.

Figure 3: Genetic Test Development and Validation and Statistical Analyses of the 2001 Anthrax Attack Evidence



Source: GAO anaysis of FBI and NAS documentation. | GAO-15-80

AMX Red Team = *Amerithrax* Red Team
CDC = Centers for Disease Control and Prevention
FBI = Federal Bureau of Investigation.
NAS = National Academy of Sciences

Note: The two genetic tests targeting the same genetic mutation associated with the D morph are referred to in this report as D-1, developed by the Illinois Institute of Technology Research Institute (IITRI), and D-2, developed by the Midwest Research Institute (MRI).

Using submission records, investigators concluded that these 8 samples were derived from a single source—a flask identified as RMR-1029. According to the FBI, this information constituted a groundbreaking lead in the development in the investigation. It allowed the investigators to reduce drastically the number of possible suspects, because only very few individuals had ever had access to this specific flask. Armed with this new information obtained from the scientific evidence, the task force focused its investigation on researchers who had had access to the laboratory at USAMRIID where RMR-1029 was stored. In 2008, the FBI sought to conduct statistical analyses in order to determine (1) the probative value of genetic markers found in a sample, and (2) possible inferences regarding the relationships of similar samples.[18] A contractor submitted the Final Statistical Analyses Report in October 2008.

In February 2010, the FBI closed the case concluding that a scientist at USAMRIID had perpetrated the attack alone.[19] Neither the case nor the totality of the evidence, including the scientific evidence that provided the FBI with valuable leads, was brought to trial in a court of law. The alleged perpetrator of the attack died on July 29, 2008, from an overdose of over-the-counter medication.

Verification and Validation

At the start of the investigation no standards or guidelines existed for verifying and validating microbial forensic methods, including the polymerase chain reaction (PCR)-based tests that were eventually used to identify the genetic mutations in the repository samples.[20] The first contractor, Commonwealth Biotechnologies (CBI), an established forensics laboratory, had begun developing the A1 and A3 genetic tests in 2002. A CBI official stated that it had relied upon the National Institute of Standards (NIST) guidance on the validation of methods for detecting human DNA, and also on the DNA Advisory Board standard for forensics. For the remaining three contractors, the Illinois Institute of Technology Research Institute (IITRI), Midwest Research Institute (MRI), and The Institute for Genomic Research (TIGR), the FBI provided "Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work"—guidelines that the FBI's Scientific Working Group on Microbial Genetics and Forensics (SWGMGF) developed and published in October

[18] Probative value refers to evidence that helps prove a fact or an issue.

[19] U.S. Department of Justice, *Amerithrax Investigative Summary* (Washington, D.C., February 19, 2010). http://www.justice.gov/archive/amerithrax/docs/amx-investigative-summary.pdf (accessed September 30, 2014).

[20] PCR is an enzymatic process by which a specific region of DNA is replicated, or amplified, during repetitive cycles to yield many copies of a particular sequence. The genetic tests all used real-time quantitative PCR.

2003.[21] These guidelines defined validation as a process by which a test procedure is evaluated to determine its efficacy and reliability for analysis.[22] Verifying and validating a test method provides a level of confidence in the ability of the test (the measurement tool) to accurately identify the properties of interest in samples that are to be analyzed.

Verification, confirms by objective evidence, from laboratory experiments, that the given test meets the user's specific requirements, such as criteria for accuracy. If the verification testing were not to produce consistent results, then the scientist or the laboratory would have to return to the optimization phase to further refine the method and materials and then revise the test's standard operating procedure (SOP) accordingly. Verification of the acceptance criteria must include repeated testing to account for measurement uncertainty, and confidence in performance statistics should be reported.[23] Depending on the intended use of a test,

[21] SWGMGF (Scientific Working Group on Microbial Genetics and Forensics), "Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work," FBI Laboratory, Quantico, Virginia, June 20, 2003, in *Forensic Science Communications* 5:4 (October 2003). http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/archives.

[22] According to SWGMGF, validation has three parts: (1) the developmental validation of a new method determines sensitivity, specificity, bias, precision, reproducibility, false positives, and false negatives of the test and determines appropriate controls (any reference database that is to be used is to be documented). (2) Preliminary validation is the acquisition of limited data to enable an evaluation of a method used to support an investigation of a biocrime or bioterrorism event. If the results are to be used for other than investigative support, then a panel of peer experts, external to the laboratory, should assess the utility of the method and define the limits of interpretation and conclusions drawn. (3) Internal validation, which should be performed and documented by the laboratory, involves testing with known samples, documentation of the test's reproducibility and precision, and a definition of the reportable ranges of the methods using controls. Before a new procedure is introduced into sample analysis, the analyst or examination team should successfully complete a qualifying test for it.

[23] Sources contributing to uncertainty include, but are not necessarily limited to, the reference standards and reference materials, methods and equipment, environmental conditions, properties and condition of the item being tested or calibrated, and the operator. In addition, statistical measures of confidence allow for the quantification of uncertainty caused by random (stochastic) errors in the performance of a genetic test. For example, one positive test result on a positive sample leads to an estimate of 100 percent sensitivity, but it provides no measure of confidence that test sensitivity actually exceeds a target of 90 percent. A misread test readout could have mistakenly caused the positive test result. Only repeated testing of positive samples can overcome random errors and ensure that test sensitivity satisfies such an acceptance criterion. If 91 or more of 100 positive samples test positive, then the genetic test is verified as exceeding 90 percent sensitivity.

sensitivity, specificity, limit of detection (LOD), reproducibility, bias, and precision may all be measures of performance (performance statistics) that should be evaluated. The type of test (qualitative, quantitative, or semi-quantitative) may also determine which of these performance parameters is to be evaluated. The testing protocol and materials, including quantities that optimize test performance are recorded in a test's SOP.

Validation confirms by examination, from laboratory experiments, and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled. Successful validation offers some assurance that a given genetic test is sufficiently robust to provide reproducible results, regardless of the practitioner, agency, contractor, or laboratory applying it to a sample. Validation is frequently used to connote confidence, but it may also be thought of as defining the limitations of a method. Studies are conducted that enable the estimation of the limits of the procedures and the measurements of the test. To the extent possible, the validation of a method should mimic "real world" conditions. The limits of the method must be known, demonstrated and documented. In essence, validation measures the uncertainty in the test output.

Reviews of the Strength of the Scientific Evidence

In 2007 the FBI convened a team of scientists to review selected scientific methods used in the case. Referred to as the AMX Red Team, it was asked to assess whether the science used was sound and to consider what additional tests might be performed to benefit the investigation. The team, finding no shortfalls or deficiencies in the basic methodologies it reviewed, concluded that the "genetic signatures correlating with specific morphs were valid tools for eliminating those repository samples not closely related to the spores used in the attack." However, the team also stated that the extent of research and development of the genetic tests at the date of its review was insufficient to determine whether the presence or absence of one or several of the morphs in a sample was associated with the evidence, was merely characteristic of normal culture practices, or possibly was affected by the genetic tests' sensitivity of detection. The team recommended additional studies to characterize the genetic markers as a function of growth conditions, including the influence of growth time, growth media, and temperature. It also recommended additional evaluation of the sensitivity of detection of each genetic test to ensure a reliable interpretation of analyses.

In 2008, the FBI asked the National Research Council (NRC) of the National Academy of Sciences (NAS) to review the scientific approaches it had used to support its conclusions. In 2011, the NAS issued its report, concluding that “it is not possible to reach a definitive conclusion about the origins of the *B. anthracis* in the mailing based on the available scientific evidence alone.”[24] Additionally, the report included the following findings related to the development and validation of the genetic tests:

> “Specific molecular assays were developed for some of the *B. anthracis* Ames genotypes (those designated A1, A3, D, and E) found in the letters. These assays provided a useful approach for assessing possible relationships among the populations of *B. anthracis* spores in the letters and samples subsequently collected for the [FBI Repository of Ames Samples] FBIR. . . . However, more could have been done to determine the performance characteristics of these genetic tests. In addition, the assays did not measure the relative abundance of the variant morphotype mutations, which might have been valuable and could be important in future investigations. . . .”
>
> “The development and validation of the variant morphotype mutation assays took a long time and slowed the investigation. The committee recognizes that the genomic science used to analyze the forensic markers identified in the colony morphotypes was a large-scale endeavor and required the application of emerging science and technology. Although the committee lauds and supports the effort dedicated to the development of well-validated assays and procedures, looking toward the future, these processes need to be more efficient.” [25]

Additionally, the NAS report included the following findings related to the statistical approach taken to quantify the significance of finding the genetic markers in a small number of repository samples:

> “The results of the genetic analyses of the repository samples were consistent with the finding that the spores in the attack

[24] National Research Council, *Review of the Scientific Approaches Used during the FBI’s Investigation of the 2001 Anthrax Letters* (Washington, D.C.: National Academies Press, Feb. 2011), p. 144. http://www.nap.edu/catalog.php?record_id=13098.

[25] National Research Council, *Review,* pp. 5–6.

> letters were derived from RMR-1029, but the analyses did not definitively demonstrate such a relationship."
>
> . . . . .
>
> "Some of the mutations identified in the spores of the attack letters and detected in RMR-1029 might have arisen by parallel evolution rather than by derivation from RMR-1029. This possible explanation of genetic similarity between spores in the letters and in RMR-1029 was not rigorously explored during the course of the investigation, further complicating the interpretation of the apparent association between the *B. anthracis* genotypes discovered in the attack letters and those found in RMR-1029."[26]

Genetic Tests Generally Were Verified and Validated but Lack of a Validation Framework Limited Statistical Confidence for Interpreting Results

We found that the genetic tests used to screen the FBI's repository of *B. anthracis* samples demonstrated through the verification and validation testing that they generally met the FBI's minimum validation requirements. However, the FBI's validation procedure did not require and the tests did not demonstrate a level of statistical confidence for interpreting the validation results. Also, tests conducted after validation—although not required by the quality assurance guidelines provided to the contractors—yielded valuable information on the performance characteristics of the genetic tests. Therefore, by not having a comprehensive validation approach, or framework, that sets out consistent steps for achieving minimum performance standards, and includes an assessment and measurement of the uncertainty in the test performance (see table 1 for the phases of a validation framework), the FBI cannot have statistical confidence in its validation test results. Knowledge of uncertainty is essential for subsequent statistical analysis that can provide quantitative measures of confidence in conclusions drawn from tests applied to forensic samples. According to DHS, it now

[26] National Research Council, *Review,* pp. 145 and 146.

validates methods and tests used to support FBI investigations and has an established ISO-accredited program.[27]

Three Phases in a Validation Framework for Genetic Test Development

In our review of scientific literature and agency and industry standards along with guidelines regarding the verification and validation of methods for analyzing both microbial and human DNA, we found that terminology and the extent of verification and validation differed across industries. However, we identified three distinct phases in genetic test development: (1) optimization, (2) verification testing, and (3) validation testing. While the literature and various validation standards and guidelines that we reviewed identify the specific types of tasks for each phase, we found that a clear boundary does not always exist between the first two phases. That is, optimization and verification are sometimes treated as a single continuous process. Further, we found that verification and validation are sometimes used interchangeably to describe the same process. Thus, the process could combine either optimization and verification or optimization and validation. Nevertheless, what is important is that the approach, or framework, to test development generally includes these phases and the associated key tasks, as shown in table 1.

Table 1: Phases and Key Tasks in Genetic Test Development

Phase	Examples of key tasks
1: Optimize the performance of a laboratory-developed method[a]	• Specify the scope, purpose and intended application of a method (e.g., genetic test) and the user's requirements • Determine appropriate controls (e.g., positive, negative) • Evaluate relevant performance parameters for a given method during this step, including sensitivity, limit of detection (LOD), specificity, precision, accuracy, reproducibility, and error rates (e.g., false positives, false negatives) • Develop a standard operating procedure (SOP) to include instructions on how to apply the test to a sample (scope, purpose, procedure, methods and materials to be tested, etc.) and criteria for interpreting the results the method generates (positive, negative)

[27] ISO standards, published by the International Organization for Standardization, headquartered in Geneva, give world-class specifications for products, services, and systems to ensure quality, safety and efficiency. Covering almost every industry from technology to food safety to agriculture and health care, they are instrumental in facilitating international trade. See http://www.iso.org/iso/home/standards.htm.

Phase	Examples of key tasks
2. Verify the performance of a laboratory-developed method	• Develop acceptance criteria (e.g., sensitivity, specificity) given the scope, purpose and intended application of the method • Develop and document a verification plan using the SOP and evaluate the verification test results against the acceptance criteria (e.g., accuracy, sensitivity, specificity, precision) to determine whether the method can repeatedly give the expected result • Use appropriate controls (e.g., positive, negative, internal, external) • Include known samples representative of those on which the method is to be used and relevant to the user's requirements • Draft an SOP that describes specific steps to follow when applying the test to the sample and criteria for interpreting the results generated by the method • Ensure that personnel are qualified to perform the method's SOP • Calculate uncertainties of measurement and characterize the method's limitations
3. Independent third-party validation of the previously verified method[b]	• Given the scope, purpose and intended application of the method, and the user's requirements, develop and document a validation test plan • Include the procedures, variables to be controlled (sample size and type) reagents, personnel, and equipment, simulate to the extent possible the conditions of the intended use of the method, and specify acceptance criteria • Include known samples, and to the extent possible case samples, that represent those on which the method is to be used and are relevant to the user's requirements • Determine that personnel are qualified to perform the method's SOP • A third party, or independent group, conducts the validation using the finalized SOP and reagent specifications; documents that the method meets the specified requirements for its intended use and acceptance criteria; use appropriate controls (e.g. positive, negative, and internal) • Calculate uncertainties of measurement and characterize the method's limitations • A subject matter expert, quality assurance officer, laboratory director, or external independent party approves the validation test results

Source: GAO analysis of scientific literature, and agency and industry guidelines for validation. | GAO-15-80

[a] The purpose of optimization is to evaluate factors (such as temperature, sample contaminants, reagent composition and the matrix) that can affect a method's performance (for example, accuracy, precision, repeatability or cross-reactivity). Conducting such experiments ensures that the most important physical, chemical, and biological parameters of that method are adjusted such that its performance characteristics are best suited for the intended application. Test performance can be evaluated by identifying known control samples and running mock tests on them.

[b] A validation test can be conducted in one laboratory or several laboratories. Conducting a validation test in more than one laboratory can provide a measure of reproducibility.

The FBI Set Minimum Performance Requirements and Relied on Contractors to Conduct Verification Testing

The FBI set limited performance requirements for the genetic tests and relied on the contractors' expertise to determine the processes they would use to develop (i.e., optimize and verify) their tests (see table 1 for types of performance parameters to be evaluated). According to the FBI, it provided minimal direction to the four contractors on how they were to develop their genetic tests in order to allow creative development. It stated that, with a few exceptions, it left the development mostly to the contractors who were experienced in developing tests. However, we found that the contractor's approaches differed in their (1) use of verification and validation guidelines, (2) steps in conducting optimization and verification testing, and (3) interpretation criteria for results generated by the genetic tests.

The FBI required that the genetic tests detect the target mutations in an overwhelming background of bacteria consisting of predominantly wild type *B. anthracis* Ames, which had been found in the evidentiary material (that is, the letters). Further, the FBI specified that sensitivity was to be demonstrated by the LOD—that is, the lowest concentration level that can be reliably detected for a qualitative and quantitative test.[28] The FBI did not require a specific calculation or value for the LOD. According to the FBI, it was looking for the presence or absence of the morphs (genetic mutations) in the repository samples and the LOD was an important factor. Three contractors developed qualitative tests; the fourth developed a semi-quantitative test.[29] In this regard, the FBI wanted to know the lowest concentration at which the genetic tests could detect the presence of a specific genetic mutation in a sample. Specificity was to be demonstrated by the detection of the target in a sample containing an overwhelming background of predominantly *B. anthracis* Ames. We found that the contractors evaluated other performance parameters at their discretion (see appendix II).

We also found that standards for the verification and validation of microbial forensics methods did not exist at the start of the investigation and were only limited after it had begun. At that time, more was known

[28] Sensitivity refers to the lowest and highest concentration of an analyte in a sample that can be quantitatively determined with an acceptable precision and accuracy. Specificity refers to the ability to assess unequivocally the analyte in the presence of components that may be expected to be present, such as impurities, degradation products, and matrix (that is, the general physical and chemical makeup of a particular sample).

[29] A qualitative test identifies the presence or absence of an analyte (what is being analyzed) such as a pathogen or toxin. Results are reported as positive or negative, or as detected or not detected. Qualitative tests may be useful for initial screening. A quantitative test provides numeric information on the amount of analyte in the sample relative to reference materials, for example. A semi-quantitative test is similar to a qualitative test in that it detects presence or absence, but it also provides a rough representation of the amount of analyte in the sample relative to a threshold—for example, \`+', \`++', or \`+++'. Thus, while the limit of detection (LOD) could be determined for both a qualitative and a quantitative test, the limit of quantitation would be determined only for a quantitative test.

about verifying and validating human DNA testing methods for forensics than about microbial forensics methods, as reflected in the revised quality assurance guidelines.[30] In addition, we found that the contractors' disparate experience and the FBI's minimal instruction to them contributed to the differences in their expectations and approaches. Most of the contractors had worked for other federal agencies whose processes differed and thus their approaches to optimizing and verifying their genetic tests differed.

While most of the contractors had developed methods for the federal government, one contractor said that each of its federal sponsors had its own processes for validation and that it followed a particular agency's processes when working with it. The contractor also stated that its own internal quality assurance guidelines were more stringent than the SWGMGF guidelines for validation. One contractor was a forensics laboratory that was familiar with analyzing human DNA samples and using associated quality assurance standards, including the DNA Advisory Board standards. Another contractor was engaged in genomic research. Finally, the FBI stated that it was more confident after the two A tests were developed; it had required the contractor for the two A tests to subject material from each polymerase chain reaction (PCR) well to genetic sequence analysis, regardless of the result (positive or negative). Further, it stated that after the first four genetic tests were developed, it had been unsure as to whether it wanted to proceed with the last one—the E test.

We found that the contractors generally conducted the tasks we identified in table 1 under the first two phases—optimization and verification—to develop the genetic tests and determine their performance, although one did not conduct a verification test. Specifically, CBI conducted an "internal

[30] SWGDAM's *Revised Validation Guidelines*, approved July 10, 2003, constituted a revision of the validation section, of FBI, Technical Working Group on DNA Analysis Methods, "Guidelines for a Quality Assurance Program for DNA Analysis," *Crime Laboratory Digest* 22:2 (1995): 21–43, https://www.ncjrs.gov/pdffiles1/Digitization/153914NCJRS.pdf.The revision was made because of "increased laboratory experience, the advent of new technologies, and the issuance of the Quality Assurance Standards for Forensic DNA Testing Laboratories by the Director of the FBI. "SWGDAM, "Revised Validation Guidelines," *Forensic Science Communications* 6:3 (July 2004), introduction, http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2004.

qualification" study that is included in the SWGMGF guidelines.[31] CBI's qualification study involved multiple experiments using internally blinded samples following an SOP to determine whether (1) the A1 and A3 genetic tests could correctly identify the targeted genetic mutations and (2) the staff involved could be considered qualified to perform the genetic tests. The first appeared to be equivalent to verification testing and the second to proficiency testing.[32] Similarly, both MRI and ITRI conducted internal verification testing and followed it with a qualification test of laboratory personnel. While the distinction between, interpretation of, and expectations for the verification and any qualification testing were not always explicit in the documentation we were provided, we found that they were both intended to precede the validation testing. However, TIGR—the last contractor to develop its genetic test—did not conduct the equivalent of either a verification or a qualification study.

The FBI indicated that it believed that the contractors had conducted verification testing but acknowledged that it was possible that one had not been conducted for the last genetic test that was developed. Thus, the verification testing was not consistent for all the tests—with one relying solely on the validation testing to determine whether it met the FBI's requirements and also was fit for use on the repository samples.

We also found that there was no clear rationale for the lack of complementary interpretation criteria for the results generated by the two genetic tests that targeted the D mutation, which proved problematic during the repository screening, after verification and validation had been completed. Each contractor independently developed interpretation criteria for positive, negative and inconclusive results through laboratory experimentation, which when defined became part of its SOP. Initially, for the A1 and A3 tests interpretation criteria were for a positive or negative

[31] According to the SWGMGF guidelines, a qualifying test measures an individual's proficiency in both technical skills and knowledge and is to be administered before assuming independent work. However, others define it as an experimental protocol that demonstrates that an accepted method will provide meaningful data for the specific conditions, matrix, and samples that the procedure is intended for.

[32] All the contractors' documentation stated that they conducted proficiency testing. A proficiency test is a quality assurance measure for monitoring performance and identifying areas where improvements may be needed. Proficiency test samples are materials whose identity, type, or values have been characterized and are used to assess the performance of a laboratory or an individual.

result only.[33] For the A1 and A3 tests, validation results were reported as the number of correct positive and negative results for the FBI-provided samples, and excluded blind samples for which, at this stage of the investigation, it was not yet known whether they contained the targeted genetic mutations.[34] For the D-1 and D-2 tests and the one E test, results were reported as the number of correct positive and negative results, detection limit, false positive rate, and inconclusive rate.

Criteria for an inconclusive result included several types of occurrences that varied by the particular genetic test.[35] The FBI stated that it reviewed the interpretation criteria for each genetic test. However, after the repository screening, disparate interpretation criteria for the D-1 and D-2 genetic tests determined the samples that had usable test results. Ultimately, contradictory interpretations of the D-1 and D-2 test results were a reason for eliminating the results of the D-1 genetic test's screening of the repository samples; thus, the D-1 results were not part of the final statistical analyses, according to the NAS report.[36] This issue did not surface during the validation of the genetic tests.

[33] These two genetic tests differed from the others in that they had a confirmatory genome sequencing step. The interpretation criteria we were provided included the results generated by the PCR as well as those from the confirmatory sequencing step. However, the validation test results were reported as either a positive or a negative. An option for an inconclusive result was available during the repository screening.

[34] Known or control samples are test materials whose identity, type, or values have been established (for example, blind samples, negative and positive controls). See "SWGMGF Quality Assurance Guidelines."

[35] Results could be inconclusive results when replicate samples generated either positive and negative results, or when a PCR reaction failed. An inconclusive result was confirmed by retesting the sample. If the retest still resulted in an inconclusive result, the final result was confirmed as an inconclusive.

[36] An FBI contractor analyzed the data for the repository samples that provided definitive results for the presence of all four morphs. However, for morph D genetic tests, the IITRI data (D-1) were dropped because their categories of results differed from those of MRI (D-2), positive, negative, and inconclusive, and included a "no growth" category outside their inconclusive results, so the final analysis was based on the MRI morph D-2 data alone. See National Research Council Review.

Validation Test Results Met the FBI Requirements but Did Not Demonstrate a Level of Statistical Confidence

We found that (1) the genetic tests used to screen the FBI's repository of *B. anthracis* samples met the FBI's validation requirements, (2) the validation tests were not required to and did not demonstrate a level of statistical confidence for interpreting the validation test results, and (3) some information on the sensitivity and specificity of the genetic tests was not characterized until after validation (postvalidation testing). As a result, the performance characteristics of the genetic tests were not fully understood when they were applied to the repository samples and more could have been done to strengthen the quality of the data and ultimately the validation results.

The validation test results showed that the genetic tests met the FBI's requirements in that they were able to detect the targeted genetic mutation when it was present at a low level (that is, at less than 1 percent, in a validation sample containing predominantly *B.anthracis* Ames), although no measurement of statistical confidence in these results was provided. As shown in table 2, the A3 genetic test had the lowest LOD, at 0.001 percent, while the others ranged from 0.005 percent to 0.01 percent.[37] The validation results led the FBI to determine that there were no false positives for any of the genetic tests, and that the inconclusive rates were 0.12 percent and 0.02 percent for the D-1 and D-2 tests, respectively. During validation, inconclusive rates could not all be computed for all the genetic tests.[38]

[37] The reported LODs, a measure of the sensitivity, were established by contractor experiments or the validation test results.

[38] Although genetic tests A1 and A3 allow inconclusive results, for the validation no test results were reported as inconclusive, and for genetic test E, none of the validation test results were reported as inconclusive, so an inconclusive rate was not calculated for the A1, A3, and E tests.

Table 2: The Limits of Detection for the Five Genetic Tests Used to Screen the FBI Repository

Genetic test	% LOD
1. A1	**0.005**
2. A3	**0.001**
3. D-1	**0.005**
4. D-2	**0.01**
5. E	**0.01**

Source: FBI data. | GAO-15-80

Note: Commonwealth Biotechnologies developed the A1 and A3 tests. The Illinois Institute of Technology Research Institute developed one of two D tests (D-1), and the Midwest Research Institute, the other (D-2). The Institute for Genomic Research developed the E test.

Since the FBI's requirements stated that the LOD was to be used as a measure of sensitivity, it was an important measure of the performance of a given genetic test. LOD provides interpretations of results generated by a genetic test with the known limitations of such data, but it is a difficult quantity to estimate reliably.[39] Our review of existing and current guidelines for validation suggests that using an appropriate estimate of LOD does provide a reliable measurement of sensitivity, but LOD estimates, like any performance statistic, should be reported with some measure of confidence. For example, the Environmental Protection Agency defines the method detection limit as the "minimum concentration of substance that can be measured and reported with 99-percent confidence that the analyte [what is being detected] concentration is greater than zero."[40] If test sensitivity is an important performance criterion, then both verification and validation procedures for a genetic test should report LOD, along with a measure of confidence. However, the LODs for the genetic tests the four contractors performed neither required a confidence measure nor determined one by using statistical measurements of confidence.

[39] Shuguang Huang, Tianhua Wang, and Min Yang, "The Evaluation of Statistical Methods for Estimating the Lower Limit of Detection," *Assay and Drug Development Technologies* (January/February 2013): 35–43.

[40] See 40 CFR Part 136, Appendix B to Part 136—Definition and Procedure for the Determination of the Method Detection Limit-Revision 1.11.

Validation testing should to the extent possible simulate the conditions of the intended use of a given genetic test, using known case samples (see tasks under the third phase in the validation framework in table 1). Calculating uncertainties of measurement is also an important task. All steps in validation testing, such as sample collection, sample preparation, transport, storage and analysis can introduce stochasticity and increase uncertainty in the test results. Most such stochasticities (and others) will also affect the testing of repository samples.[41] These additional uncertainties can be measured and understood using repeated (replicate) experiments including all relevant steps (from collection to analysis) of samples with known concentration levels. By designing a validation study with a sufficient number of replicate samples, the FBI could have quantified the level of statistical confidence in the sensitivity and specificity of the tests.

While the SWGMGF guidelines did not require them, tests as part of validation that examined stochastic (random) effects of the process would have made it possible to draw more rigorous conclusions, with measures of confidence, regarding the test results for the repository samples.[42] In addition, we found that additional information on the sensitivity and specificity of the PCR-based genetic tests was characterized during postvalidation testing that the FBI's expert advisers recommended. Our analysis of these post-validation test results suggests that the negative rates of the genetic tests were high for samples that could be expected to contain the genetic mutations when using the sample collection and processing methods as required for the repository samples and that there were stochastic (random) effects in the repository screening process.

The FBI's expert advisers had reviewed information on the genetic tests, including the validation data, and recommended additional tests to better understand some of the uncertainties in the preparation and analysis of a repository sample. The FBI and its advisers had recognized that because each contractor had developed its genetic test to detect a specific genetic

[41] Stochastic is synonymous with "random." Greek in origin, the word means "pertaining to chance." http://mathworld.wolfram.com/Stochastic.html (accessed October 17, 2014).

[42] The SWGDAM guidelines do suggest such tests for PCR-based methods, which is important when samples contain low concentrations of the target to be detected. Sampling fluctuations can occur in PCR-based tests. According to the 2004 SWGDAM guidelines, for PCR-based assays, validation studies must address stochastic effects and sensitivity levels.

sequence, growth conditions would vary slightly. However, according to the FBI, the purpose of the additional testing was to determine whether stochastic sampling error had been introduced into the repository preparation process as instructed in the subpoena.[43] Therefore, the postvalidation tests were to determine whether the procedures by which the repository samples were processed could affect the accuracy of the interpretation of the data.[44] The postvalidation tests were conducted in August and September 2007 under conditions that closely mimicked the intended use for each of the genetic tests. According to the FBI, the screening of the repository samples with the genetic tests was about three-fourths complete when this testing took place.

Specifically, in the postvalidation tests, the contractors applied their genetic tests to replicate samples derived directly from some of the evidentiary material—including flask RMR-1029.[45] The results revealed that the genetic tests did not always detect the genetic mutations in samples that had been derived directly from the evidence and thus were expected to contain all four mutations—a best-case scenario.

Our evaluation of measures of the sensitivity and specificity of the genetic tests revealed differences between the validation and postvalidation test results. Regarding sensitivity, under the assumption that undiluted samples from flask RMR-1029 are positive for all four genetic mutations (supported by the preponderance of genetic and non-genetic data), we can estimate the negative rate as the frequency of negative results in replicate tests of undiluted samples from RMR-1029.

[43] According to the FBI, it and its advisers recognized that the subpoena instructions might not have been sufficiently clear to ensure that an adequate amount of sample was used to create the FBI repository exemplars. Variations in the samples, such as sample density, percent viability, and the amount of sample taken could have resulted in the collection of insufficient sample for a reliable analysis of the original material.

[44] In addition, many explanations are possible for variation in quantitative PCR results, including differences in temperature, concentration, and stochastic (random) variation. Precision in PCR typically varies with concentration.

[45] To simulate submissions to the repository, 30 samples were taken from flask RMR-1029 (expected to contain all the morphs), the original 1981 *B. anthracis* Ames, and another sample. The contractors used their respective genetic tests to analyze these samples—which were in both undiluted and diluted form.

Validation testing showed that for those results expected to be positive, no negative results were observed at or above the LOD for any of the genetic tests.[46] However, in the postvalidation testing, the negative rates were generally high. As shown in table 3, the negative rates for the postvalidation tests ranged from 0 percent to 43 percent for the undiluted samples from flask RMR-1029. (Appendix III breaks down the results of the replicate testing for each genetic test.)

Table 3: Sensitivity Results for Five Postvalidation Tests on Undiluted Samples from Flask RMR-1029

	Number		Sensitivity	
Genetic test	Replications from flask (positive samples)	Positive samples detected	Nonpositive results[a]	Estimated % negative rate[a]
A1	30	17	13	43.3
A3	30	29	1	3.3
D-1	30	23	7	23.3
D-2	30	24	6	20.0
E	30	30	0	0

Source: FBI, sensitivity statistics derived from 30 replicate samples selected from RMR-1029 using sample selection methods similar to the samples submitted to the FBI repository. | GAO-15-80

[a] Includes negative and inconclusive results as nonpositive results. The estimated negative rate is the number of non-positive results divided by the number of replications.

The NAS report stated that the FBI did not address false negative results and inconclusive results, and it was concerned about the restriction of the statistical analyses to the repository samples that had no inconclusive or

[46] For the D-1, D-2, and E genetic tests, no negative results were observed at or above the reported LOD in the validation testing. For the A1 and A3 tests in the documentation we were provided, the validation test results were not tied to the reported LOD and no information was provided on the concentration of targeted genetic mutation and wild type in the validation samples. Both detected all 6 positive samples. For the 10 unknown samples (not known whether they contained the target genetic mutation) it analyzed in the test, the A3 test generated positive results for all. For 7 unknown samples it analyzed, the A1 test generated positive results for 5—with 2 negative results. The FBI calculated an inconclusive rate for the D-1 and D-2 tests at 0.02 percent and 0.12 percent, respectively. However, inconclusive rates could not be computed for the A1, A3, and E genetic tests, as discussed earlier.

variant results.[47] Of the two genetic tests that targeted the D mutation, the results of only D-2 were used in the FBI's analysis of the repository screening—that is, the analysis was restricted to the 947 samples that contained no inconclusive or variant results, which resulted in the exclusion of 112 samples from the analysis. Thus, the knowledge about sensitivity and specificity obtained by the replicate testing, as well as ensuring that these two genetic tests' interpretation criteria were complementary, would have been more useful if it had been completed in the validation process.

Regarding measures of specificity, the effect of the repeated analysis of the undiluted nonpositive samples during the postvalidation testing showed evidence of a nonzero false positive rate for the D-2 genetic test. As shown in table 4, the 3.3 percent false positive rate for the D-2 genetic test demonstrates the likelihood of a random effect in the postvalidation tests that was not apparent from the validation results.

[47] In commenting on a draft of the report, the FBI stated that we are using the term "false negative" incorrectly, suggesting that the genetic tests were inadequate and prone to producing false negative results. It stated that the negative results that we refer to as false negatives are the result of stochastic sampling error. The FBI defined a false negative as "when a positive sample fails to detect the analyte of interest when the sample is known to contain the analyte of interest at detectable levels." We understand that the purpose of the postvalidation testing was to determine if the sampling instructions provided in the subpoena introduced sampling error, as the FBI stated in its comments. However, in our report, we are focusing on the process as a whole—from sampling to analysis. Error can occur at any point in this process. The evidence we present shows that results from some samples in the postvalidation tests that were expected to be positive were negative. While these results could be due to stochastic error, they are still unexpected results. Thus, the process can cause variation in a test result.

Table 4: Specificity Results for Five Postvalidation Tests on Undiluted Samples of Wild Type *B. anthracis*

	Number			Specificity	
Genetic test	Replications from flask (negative samples)	Nonpositive samples detected[a]	Positive results	Estimated % false positive rate[b]	Estimated % false positive rate from validation tests[c]
A1	30	30	0	0	0
A3	30	30	0	0	0
D-1	30	30	0	0	0
D-2	30	29	1	3.3	0
E	30	30	0	0	0

Source: FBI, specificity statistics derived from 30 replicate samples selected from the wild type (no genetic mutations) material using sample selection methods similar to the samples submitted to the FBI repository. | GAO-15-80

[a] Nonpositive results include both negative and inconclusive results.

[b] The estimated false positive rate is the number of positive results divided by the number of replications.

[c] No false positives were observed at or above the limits of detection for the genetic tests on the basis of the validation test results.

Although not a requirement at the time, repeated testing—such as that conducted postvalidation—would have provided additional information on the performance of the genetic tests. We recognize that neither the FBI nor the SWGMGF guidelines required contractors to conduct replicate tests of case samples to identify the stochastic (or random) effects of the genetic tests when they were used under realistic test conditions to further evaluate the genetic tests' sensitivity—an important step in validating PCR-based genetic tests. In contrast, the SWGDAM guidelines suggested using experiments to determine the sensitivity of real-time PCR-based tests as a part of validation.[48] Importantly, while the LOD is a critical performance indicator for a genetic test, LOD calculations do not account for the data that PCR-based tests sometimes generate but that are not typical.[49] The FBI also stated that during the development of the

[48] The SWGDAM guidelines suggested additional testing to identify a method's stochastic (or random) effects when testing for sensitivity. For example, the 2004 guidelines stated that, among other things, for PCR-based procedures, "The laboratory must conduct studies that ensure the reliability and integrity of results. For PCR-based assays, studies must address stochastic effects and sensitivity levels." SWGDAM," http://www2.fbi.gov/hq/lab/fsc/backissu/july2004/standards/2004_03_standards02 (accessed September 30, 2014).

[49] According to an expert we consulted, the LOD can only be estimated from engineered samples with known analyte concentrations. Thus, LOD as a measure of sensitivity may not correctly estimate the sensitivity of the test applied in more realistic scenarios.

genetic tests it was concerned that stochastic effects might be a problem, stating, for example, that it had discussed its concerns with the contractors about evidence growth steps and the possible stochastic effects, that is, in the context of the growth rates of the wild type cells (*B. anthracis Ames*) versus the morph cells in culturing, among other things. The postvalidation tests were able to estimate valuable performance statistics of the genetic tests and under more realistic testing conditions than the original validation tests.

More extensive validation testing could have reduced uncertainties in the testing procedure. For example, the sensitivity of a given genetic test relies on the sampling procedures, the rarity of the targeted genetic mutation in a sample, and other factors that vary by genetic test. Incorporating these types of tests into the validation would have resulted in more information on the uncertainties inherent in the use of the genetic tests and would have been a way to simulate the conditions of their intended use. Future validation efforts would be strengthened by including experiments designed to identify and eliminate likely uncertainties in test performance.

The differences we have highlighted regarding the contractors' approaches to verification and validation indicate that the use of a comprehensive validation framework could help ensure greater consistency. Such a framework would need to specify the defined level of statistical confidence to be calculated for the interpretation of validation results before they are applied to evidentiary samples. Minimally, the statistical confidence achievable in each test should be estimated during validation.

The development of such a framework could be facilitated by DHS's National Bioforensics Analysis Center (NBFAC), which validates tests used to support FBI bioforensic investigations. According to DHS, NBFAC will take steps to ensure that the results it generates will meet *Daubert* standards for "appropriate validation" and third party review and will thus meet admissibility requirements for evidence in federal court proceedings.[50] NBFAC—an ISO 17025 accredited forensic laboratory—is experienced in working with multiple outside laboratories to verify and validate their methods. It has an established ISO 17025 accredited process.[51]

The combination of limited communication among the contractors, varied timing in the validation efforts, uncertainties the FBI faced as the investigation unfolded, and increasing knowledge about the repository samples made it clear, with hindsight, that the contractors' verification and validation approaches were likely to differ. Thus, in the future, standardizing the approach to verification and validation testing—by the means of a validation framework—would be more efficient, especially in clearly communicating expectations to multiple contractors.

DHS Validates FBI's Microbial Forensics Methods That Can Support FBI Investigations

In contrast to 2001, DHS's NBFAC validates assays (or tests) that can be used to support FBI bioforensic attribution investigations. Generally, the NBFAC validation process involves the evaluation of methods transferred from others, such as DOD and academic laboratories, and sometimes the development of a new method. For forensic tests, NBFAC and the FBI are provided with a "validation package" for each test that encompasses data on testing previously conducted during the development stage or

[50] Under the Federal Rules of Evidence, Rule 702, an expert witness is considered qualified to testify if, among other things, the testimony is the product of reliable principles and methods. The 1993 Supreme Court case, *Daubert v Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993), significantly changed the admissibility of scientific evidence for federal trial courts. The *Daubert* case listed factors for judges to use in assessing the reliability of scientific expert testimony, including (1) whether the expert's technique or theory can be or has been tested, (2) whether the technique or theory has been subject to peer review, (3) the known or potential rate of error of the technique or theory when applied, (4) the existence and maintenance of standards and controls, and (5) whether the technique or theory has been generally accepted by a relevant scientific community. We refer to these factors as the *Daubert* standards.

[51] NBFAC is accredited through its third-party accrediting body, the American Association for Laboratory Accreditation (A2LA) and it has an ISO 17025 quality management program. A laboratory's accreditation to the ISO 17025 standard ensures that it is technically competent to provide accurate and reliable results. Technical competence requires qualified staff, properly calibrated and maintained equipment, appropriate and validated test methods and procedures, traceability to national standards, accurate recording and reporting procedures, suitable test facilities, procedures for the proper handling of test items, and quality assurance procedures. Once a laboratory is accredited, competence is ensured by periodic evaluations, proficiency test programs, and external assessments.

before the transfer to the laboratory.[52] According to DHS, developers have to provide information on the performance parameters (e.g., accuracy, LOD, precision) that they have previously verified. Next, NBFAC conducts its own test, evaluation, and validation of the transferred method. When evidence stemming from the use of validated methods is needed as evidence in court, it must be defensible by meeting evidentiary standards.

Questions may be raised in court about the standards used for the validation of such methods. Results generated by forensic methods, including microbial forensics, must meet a high standard. According to NBFAC, to ensure that results generated by a validated test will meet *Daubert* standards for "appropriate validation," the deliverables from the Bioforensics R&D Program include SOPs for the methodologies and technical and peer-reviewed published reports. Also, quality project performance plans are required of researchers, who must define method performance parameters to provide a baseline for verification and further validation if required by law enforcement.[53] NBFAC's ISO 17025 accreditation of tests requires the demonstration of previously described method parameters in NBFAC laboratories with trained staff followed by a third-party review of the supporting data, procedures, equipment, and staff training that supports ISO 17025 accreditation.

[52] According to NBFAC, tests established within NBFAC have been developed within the government biodefense community, DHS's Bioforensic R&D Program, academia, and the commercial sector and their performance is verified and validated within the NBFAC ISO 17025 accreditation program.

[53] In this context, a method that has been validated elsewhere and that is transferred would be evaluated to ensure that the NBFAC successfully used the method as intended in the NBFAC laboratory. Performance parameters include accuracy, precision, specificity, selectivity, LOD, limit of quantitation, linearity, ruggedness, and robustness.

Characteristics of a Statistical Framework That Would Strengthen the Significance of Microbial Forensic Evidence in Future Investigations

We identified six characteristics of a statistical framework that would strengthen the significance of microbial forensic evidence.[54] When we compared the FBI's statistical approach to these six characteristics, we found that three could be improved to strengthen the significance of its evidence for future investigations.[55] That is, the FBI (1) could do more to understand the methods and conditions that give rise to the chosen genetic markers, (2) institute more rigorous controls over sample identification and collection, and (3) include measures of uncertainty when interpreting the results. We found that the FBI has taken some steps to include such expertise in future investigations by building formal forensic statistical expertise both internally and externally.

The Six Characteristics of a Statistical Framework

Although not always possible, an important goal of a microbial forensic investigation is to generate meaningful comparative analyses of evidentiary samples and suspect samples to establish their relatedness or to exclude suspect samples from an investigation. Statistically meaningful comparative analyses can allow the use of statistical inferences relating to the process to produce the sample, the provenance of a sample, or the relatedness of samples.[56] The significance of such statistical inference relies on the analyst's ability to quantify both the confidence in test results and the frequency with which results match. Confidence, in this context, refers to the level of reliability and accuracy investigators assign to the test results obtained from the measurement tools used to identify the properties of interest in the samples. The frequency of the sample properties' presence, or generation in a relevant population of possible sources, is a measure of how common or rare the properties are and provides context to the probative value of the evidence. According to a 2009 NRC report, a statistical framework is needed to quantify the

[54] A statistical framework allows for statistically meaningful comparative analyses; it is a set of concepts and organizing principles that support the compilation and presentation of a set of statistics.

[55] The statistical approach in analyzing repository samples included identifying the strain of *B. anthracis*, collecting the repository, conducting statistical analyses of the repository samples, and presenting and interpreting the results of the genetic test.

[56] Such analyses include the computation of association statistics, probability estimates, confidence intervals, and statistical tests of significance related to specific hypotheses.

probative value of forensic evidence in terms of the frequency of that evidence in a population.[57]

Formulating an appropriate statistical framework that is adequate for all microbial forensic investigations is not feasible because the diversity of many potential pathogens is unknown or, at best, difficult to describe. For this reason, frameworks must be adapted to the specific circumstances of each case. As shown in table 5, our review of scientific literature in forensic science, statistics, epidemiology, and population genetics identified the six general characteristics that a framework needs for statistically meaningful comparative analyses of the attack material to repository samples for the specific set of circumstances of the FBI's investigation.

Table 5: Six General Characteristics of a Framework for Statistically Comparing Attack Material to Repository Samples

Characteristic	Definition
1	The genetic signature used to determine a match or exclusion should be clearly defined and understood
2	A relevant source population should be clearly defined and understood
3	A database that accurately and completely represents the genetics of the relevant source population should be created
4	The limitations of measurement tools (or assays) should be known
5	The statistical methods should be appropriate for the data and should properly account for the mode of inheritance of the genetic markers
6	The interpretation of results should include quantifications of uncertainty

Source: GAO. | GAO-15-80

First, a definition of what constitutes a matching type should be clearly established. A genetic signature, or a set of genetic markers, can be chosen to establish a genetic type (or genotype) that is used to differentiate the samples. The genetic signature should be sufficient to identify the target of interest at the resolution needed for an investigation. In this case, the target of interest was the *B. anthracis* Ames strain, capable of producing spores with a set of specific genetic markers linked to morphs observed after a prolonged period of growth. The requisite resolution was the ability to differentiate among the individual stocks (or

[57] National Research Council, *Strengthening Forensic Science in the United States*: *A Path Forward* (Washington, D.C.: National Academies Press, 2009), p. 189. http://www.nap.edu/openbook.php?record_id=12589.

collections of organisms) of *B. anthracis*. Determining that two or more samples have a matching type must take into account the source of the organisms (for example, nature or the laboratory), the stability of genetic markers, storage conditions, and conditions giving rise to the markers. Specific growth or environmental conditions may selectively advantage or disadvantage mutations and affect the stability of genetic markers. Therefore, if the significance of a matching genetic signature is to be understood, the genetic markers should be well characterized, and the conditions giving rise to the presence of markers in a sample should be understood.

Second, once the genetic signature has been established and a match has been clearly defined, it is then necessary to identify and define the population of relevant sources that may have the genetic signature in order to understand how common or rare the genetic signature is. This relevant source population is critical in identifying the probative value of any match or nonmatch between samples. In a criminal investigation, a relevant source population may be considered the population of suspects, and it should be defined as specifically as possible to identify the smallest population related to the evidentiary material. The definition of the relevant source population should be based on the population related to characteristics of the evidence and not on characteristics of a suspected source. The relevant source population in this case is all stocks of *B. anthracis* that could have been used to grow the material used in the attack letters.

In defining the source population, the structure of the relevant source population of bacteria should be understood. When a population is divided into subgroups that do not mix freely, that population is said to have structure. In this case, the relevant reference population of stocks of *B. anthracis* was highly structured among the laboratories included in the investigation. The lack of independence between stocks in a structured population affects inferences about the evidentiary material and its most and least likely sources.

Third, in order to quantify or estimate how common or rare a genetic signature is in the relevant source population, a database that accurately represents the relevant source population's genetics should be created. The extent to which the database reflects the population will affect the accuracy of the match probability. The size and quality of the data in the database will affect the power of match probability, determining the potential probative power of the signature for distinguishing one source from another. A large and comprehensive database is the theoretical goal but in most cases may not be possible. However, in this case, the FBI determined that it was possible to identify all sources of the *B. anthracis* Ames strain, and it set out to create a comprehensive database. For completeness the genetic information in the database should have included samples from all sources of the *B. anthracis* Ames strain. In such cases, the database should be complete—excluding sources results in underrepresentation—and should avoid duplication (although replication can be beneficial)—unknowingly including sources more than once results in overrepresentation. Methods used to select samples from each stock should be adequate to ensure representation of the organisms within each stock. In an ideal situation, the database of genetic information should be constructed to the same quality standards as the actual evidentiary analysis. These quality standards should apply to the selection of samples from stocks to the results of the genetic tests.

Fourth, the limitations of the measurement tools used to generate the genetic information in the database should be identified. When quantitative inference is attempted, care must be taken not to overemphasize data; the limits of the methods used to generate the data should be considered. The power and limitations of microbial forensics methods need to be understood through validation. Validation frequently connotes confidence, but it may be thought of as defining the limitations of the method. This does not mean that a method must be 100 percent accurate to be useful. Studies should allow the estimation of the limits of the measurements. The limits of the methods must be demonstrated and documented for all steps in the process, including sample collection, preservation, extraction, analytical characterization, and data interpretation.

Fifth, the choice of statistical methods should be appropriate for the data and should properly account for the mode of inheritance of the genetic markers and any structure in the populations. An important aspect of computing association statistics and probability estimates is properly accounting for the mode of inheritance. Methods appropriate for computing probability estimates and statistical tests of significance differ by the mode of inheritance of the genetic markers. In organisms that reproduce asexually, such as *B. anthracis*, genetic diversity is driven by mutation processes, not by random mixing. Computing match probabilities using methods that assume independence and random mixing within populations is not appropriate because the genetic variation in such organisms is highly correlated. In organisms that reproduce asexually, the frequency of a particular genetic type in the population must be determined by direct observation. The frequency of the evidentiary genotype in a relevant source population can be based on counting the number of times the genotype is observed in a reference database. The strength of this approach is affected greatly by the genetic database and whether it has sufficiently sampled relevant populations.

Sixth and finally, the interpretation of results should include quantifications of uncertainty. It is crucial to clarify the type of question the analysis is addressing when evaluating the accuracy of a forensic analysis. Although some techniques may be too imprecise to permit the accurate identification of a specific individual, they may still provide useful and accurate information about questions and classification. The interpretation of results will be stronger with the proper use of statistical and probabilistic analyses, but the strengths and weakness of any result should be communicated. Results should indicate the uncertainty in the measurements, and studies must be conducted that enable the estimation of those values.

The FBI's Statistical Approach Could Have Been Improved for Three of the Six Characteristics of a Statistical Framework

We believe that the six general characteristics described above make up a comprehensive statistical framework that could have allowed the FBI to quantify significance and probative value of the scientific evidence collected in a statistically meaningful way and could have strengthened the evidence it collected. However, we found that at the outset of its investigation, the FBI did not have a comprehensive framework that would allow for statistically meaningful comparative analyses between samples from the attack letters and samples in the FBI repository of *B. anthracis* Ames strain. Specifically, we found that the FBI's approach to three of the six characteristics could be improved to strengthen the significance of evidence in future investigations.

The FBI's Research Did Not Provide Full Understanding of the Methods and Environmental Conditions That Give Rise to Genetic Mutations

Although the specific genetic mutations used as genetic markers to determine a match or exclusion were adequately characterized, the FBI did not conduct studies to understand the methods and environmental conditions that gave rise to the mutations. The FBI convened a team of scientists in 2007 to review the scientific methods. Finding no shortfalls or deficiencies in the basic methodologies they reviewed, they determined that the usefulness of the genetic markers was sufficient. The team also stated that the extent of research and development of the genetic tests at the date of their review was insufficient to determine whether the presence or absence of one or several of the genetic markers was associated with the evidence, was merely characteristic of normal culture practices, or possibly was affected by the sensitivity of detections of the genetic tests. The team recommended additional studies to characterize the genetic markers as a function of growth conditions, including the influence of growth time, growth media, and temperature.

In response to questions from the NAS panel about this recommendation, the FBI stated that it considered such studies academic and did not conduct the recommended research. Consequently, experimental data are missing that would have shown the frequency with which particular genetic mutations occur under growth conditions that could affect their retention or loss. In its report, NAS opined that some of the morphs used as genetic markers might have arisen independently from RMR-1029.[58] According to the report and experts we spoke with, the genetic markers might have had a selective advantage under growth conditions used for large-scale production of spores, such as in a fermenter or in a batch culture. If so, the presence of the genetic markers would be a function of the growth conditions rather than direct derivation from parent material, such as RMR-1029. This is problematic for the quantification of the rarity of the results because it is not possible to calculate the probability of two independent cultures having the same genetic markers if either was subjected to growth conditions that provide selective advantage or disadvantage.

Without the experimental data, the usefulness of the genetic markers as an identifying signature to determine a match or exclusion was not fully understood. For example, it is not known whether the genetic markers could have arisen independently. To identify repository samples that received a direct or indirect transfer from the laboratory that possessed RMR-1029 after it was created in 1997, we examined the FBI's documentation of historic transfer records of *B. anthracis* Ames strain between laboratories from 1981 through 2001. We supplemented this with information from laboratory officials and researchers we interviewed. Then, we compared the frequency of positive genetic markers in these

[58] Finding 6.3 of the NAS report states that some of the mutations identified in the spores of the attack letters and detected in RMR-1029 might have arisen by parallel evolution rather than by derivation from RMR-1029. The investigation did not rigorously explore this possible explanation of genetic similarity between spores in the letters and in RMR-1029, further complicating the interpretation of the apparent association between the *B. anthracis* genotypes discovered in the attack letters and those found in RMR-1029.

groups of samples to the 119 samples that we verified were independent of transfers from the laboratory that possessed RMR-1029.[59]

Our analysis of repository data found no evidence of independent evolution in three of the four genetic markers (A1, A3, and E). However, we found that repository samples with no direct or indirect relationship to RMR-1029 tested positive for the D genetic marker at rates similar to those of the samples that were submitted from laboratories with direct transfers from the laboratory that possessed RMR-1029.[60] As shown in table 6, the D genetic marker was detected in about 6.6 percent of the repository samples submitted from laboratories with direct transfers from the laboratory that possessed RMR-1029 compared to 6.7 percent of the samples that were independent of the laboratory that possessed RMR-1029.

Table 6: The Percentage and Frequency of Positive Results for Genetic Tests in the FBI's Repository Following the 2001 Anthrax Attack by Direct, Indirect, and Independent Transfer Path

	Transfer path		
Genetic test	**Direct**	**Indirect**	**Independent**
A1	(32/739) = 4.3%	(3/201) = 1.5%	(0/119) = 0.0%
A3	(20/739) = 2.7%	(1/201) = 0.5%	(0/119) = 0.0%
D	(49/739) = 6.6%	(19/201) = 9.5%	(8/119) = 6.7%
E	(22/739) = 3.0%	(1/201) = 0.5%	(0/119) = 0.0%

Source: GAO analysis of FBI repository data, 2014. | GAO-15-80

Note: Percentages indicate positive samples. Results for the two D tests were treated as a single positive if either the D-1 or D-2 assay showed a positive result. In this analysis all nonpositive results (inconclusive, variant, and no growth) were treated as negative results. This represents the most conservative assumption.

[59] We did not obtain information on the origin of all samples within the laboratories or the specific samples that were associated with all transfers. This analysis was based on between-laboratory transfer records; the classification of samples into the direct or indirect groups does not indicate whether a sample was descended from RMR-1029. However, samples submitted by laboratories classified in the independent group have, by definition, no relationship to RMR-1029.

[60] We identified four laboratories that received a direct transfer from the laboratory where RMR-1029 was found (direct transfer path) after RMR-1029 was created; six laboratories that received a transfer from the laboratory where material was grown that was used in the creation of RMR-1029 (indirect transfer path), and 7 laboratories that had no direct or indirect transfers (independent transfer path).

Additionally, the NAS report found that in repository samples associated with experiments conducted before the 2001 attacks, the D genetic marker was the only marker detected and it occurred in about 1 percent (3 of 296) of those samples. This provides additional evidence that the D genetic marker may have arisen independently of RMR-1029. Additional studies recommended to the FBI that it did not conduct could have provided the experimental data needed to fully understand the probative value of this genetic marker.

The FBI Established an Adequate Relevant Source Population

Because the FBI adequately identified the relevant source population as all stocks of *B. anthracis* Ames strain, it significantly reduced the number of possible sources. The NAS report found that the dominant organism in the letters was correctly and efficiently identified as the Ames strain of *B. anthracis*. The science performed on behalf of the FBI for identifying the *Bacillus* species and *B. anthracis* strain was appropriate, was properly executed, and reflected the contemporary state of the art. The correct identification of the specific strain of *B. anthracis* allowed the FBI to adequately define the relevant source population as stocks of the Ames strain in laboratories that had the Ames strain in their inventories before the attacks. This significantly reduced the number of possible sources.

The FBI Did Not Fully Ensure the Completeness and Accuracy of the Repository

We found that the FBI's effort to create a comprehensive repository containing samples from all known stocks of the Ames strain of *B. anthracis* was appropriate for assessing the rarity of the genetic markers in the relevant source population. Its adequacy, however, was affected by the incompleteness and inaccuracy in the repository. The NAS report found that the repository was not optimal for a variety of reasons. It stated, for example, that the instructions in the subpoena issued to laboratories for preparing samples were not precise enough to ensure that they would follow a consistent procedure for producing samples that would be most suitable for later comparisons.

Our analysis of FBI documents shows that FBI searches at three specific laboratories identified hundreds of additional relevant stocks that laboratories did not submit to the repository in response to the subpoena. Specifically, we found that the FBI collected about 29 percent of the 1,059 repository samples through these searches. The proportions of samples thus obtained were 34 percent, 96 percent, and 22 percent in these laboratories (see table 7).

Table 7: Number and Percentage of Repository Samples from Three Searched Laboratories Following the 2001 Anthrax Attack

		Number of samples		
Laboratory	**Total**	**From subpoena**	**From search**	**% of samples from search**
A	**684**	450	234	34%
B	**71**	3	68	96
C	**27**	21	6	22
All other	**277**	277	0	0
Total	**1,059**	**751**	**308**	**29**

Source: GAO analysis of FBI repository data, 2014. | GAO-15-80

We were unable to determine how two of the three laboratories identified and selected samples from relevant stocks in response to the subpoena, but we found that individuals at one laboratory differed in interpreting the subpoena's instructions.[61] Laboratory officials acknowledged differences in interpreting the instructions on how to identify distinct Ames strains of *B. anthracis*. Identifying the specific stocks to submit in response to the subpoena at that laboratory was left up to the principal investigator because, at that time, no one else actually working with the stocks would have understood what was in them.

FBI officials acknowledged that the interpretation of the instructions to determine what strains to submit to the repository varied across laboratories, stating that the subpoena was not as precise as it needed to be. However, they emphasized that every laboratory that submitted samples to the repository was investigated thoroughly and that, when the FBI conducted searches at the three laboratories, those investigations eliminated many laboratories from being suspects. Furthermore, FBI officials told us that the decision to conduct searches at these three laboratories was an investigative decision, not a scientific one.

The NAS report also raised concerns that the decision to remove samples with inconclusive or variant results contributed to the lack of completeness of the repository data. The report stated that a major concern was the restriction of its statistical analyses to the 947 samples

[61] We were unable to determine how two laboratories identified and selected samples from relevant stocks because of staff retirements, the attrition of key technical staff, and the absence of inventory records for the period of the investigation.

that contained no inconclusive or variant results. Notably, the report showed that 4 of the 112 samples that were disregarded for having a single inconclusive or variant result scored positive for the three remaining genetic tests.

In addition, our analysis of FBI documents shows that FBI searches contributed to inaccuracies in the repository by collecting samples from stocks that had already been submitted to the repository. We identified 14 duplicate samples from a search conducted at one laboratory in April 2004.

FBI officials stated that they were not concerned about duplicate samples in the repository because duplicate samples may have served other important investigative purposes such as verifying if two samples were related or answering other important questions related to investigative information. They also stated that additional information collected about the samples would allow them to reconcile duplicates. However, our analysis of the FBI repository data indicates that known duplicates were not removed from the repository before the statistical analysis.

As a result of these examples of incompleteness and inaccuracies in the repository, a statistically meaningful extrapolation of the statistics and frequencies derived from the repository to the relevant source population was not possible. By instituting more rigorous controls over sample identification and collection for future investigations, the FBI can improve the completeness and accuracy of a repository.

The Statistical Analyses Did Not Account for the Genetic Markers' Mode of Inheritance

The results from statistical analyses conducted in 2008 did not adequately account for the mode of inheritance of the genetic markers, and they added little probative value to the investigation. Many of the methods used for the 2008 statistical analyses inappropriately relied on the assumption of independence among the repository samples. For example, the NAS report stated that because the repository samples were not independent, the proportion of samples testing positive for all four genetic markers was not a meaningful estimate of the probability of occurrence. The FBI did not use the results of the statistical analyses and did not quantify the confidence it had in, or the probative value of, the repository results in its conclusions included in its final investigative summary. An FBI official stated that the statistical analyses were viewed from an academic standpoint and were not part of the investigation. That official also stated that the results of the statistical analyses did not contradict the conclusions of the investigation.

The FBI Did Not Include Uncertainty Measures in the Interpretation of Results

In its final investigative summary, the FBI concluded that only 8 of more than 1,000 samples tested positive for all four genetic markers, but it did not provide any measure of the confidence it had in this conclusion.

We found that the genetic tests show variability in the results on samples selected from the same stock. As we previously indicated in our assessment of the validation of the genetic tests, the additional postvalidation tests conducted in 2007 demonstrated variability in the results of the genetic tests when they were applied to samples under conditions intended to mimic their use on repository samples.

Additionally, the two genetic tests for the D marker did not always give the same result for the same sample. An analysis included in the FBI contractor's Statistical Analysis Report identified 24 repository samples for which the two genetic tests yielded opposite results from the same sample. The NAS report stated that this lack of agreement between the two genetic tests for the D mutation illustrated the differing sensitivities and specificities of the tests. This lack of agreement was also evident in the eight samples that tested positive for all four genetic markers. As shown in figure 4, our analysis of the repository data demonstrated that one of these eight samples also tested negative using the other genetic test for the D marker.

Figure 4: The Results of Genetic Tests for the Eight Samples Testing Positive for All Four Markers Following the 2001 Anthrax Attack

FBIR sample	A1	A3	D-2	D-1	E
005-016	+	+	+	+	+
044-040	+	+	+	+	+
049-004	+	+	+	+	+
049-006	+	+	+	+	+
049-008	+	+	+	+	+
049-016	+	+	+	+	+
053-070	+	+	+	-	+
054-076	+	+	+	+	+

-	Negative result
+	Positive result

Source: GAO analysis of FBI data, 2014. | GAO-15-80

Note: Includes both tests for the D marker.

Further, our analysis of duplicate samples in the repository showed differences in the results of genetic tests on samples selected from the same stock. As shown in figure 5, only 3 of the 14 duplicate samples we identified showed the same results across the five genetic tests. For example, FBI repository sample number 049-004 tested positive for all five genetic tests while a duplicate sample selected from the same stock (066-044) tested positive for only four of the five genetic tests. In another example, FBI repository sample number 049-016 tested positive for all five genetic tests while the duplicate sample (047-002) tested negative for all five genetic tests.

Figure 5: Duplicate Repository Samples with Results of Genetic Tests Performed after the 2001 Anthrax Attack

FBIR sample	A1	A3	D-2	D-1	E		A1	A3	D-2	D-1	E	FBIR sample
049 - 002	-	-	i	-	-	⇔	v	-	-	+	-	044 - 034
049 - 004	+	+	+	+	+	⇔	+	+	+	+	-	066 - 044
049 - 008	+	+	+	+	+	⇔	-	+	-	i	+	044 - 020
049 - 014	i	v	+	+	-	⇔	i	-	-	i	-	041 - 006
049 - 016	+	+	+	+	+	⇔	-	-	-	-	-	047 - 002
049 - 018	-	v	+	+	-	⇔	-	-	-	-	-	041 - 004
049 - 020	-	-	-	-	-	⇔	-	v	-	-	-	044 - 042
049 - 024	-	-	-	-	-	⇔	-	-	-	-	-	044 - 026
049 - 026	-	-	+	-	-	⇔	-	-	-	-	-	044 - 022
049 - 028	-	-	-	-	-	⇔	-	-	-	-	-	041 - 002
049 - 036	-	-	-	-	-	⇔	-	-	-	-	-	044 - 024
049 - 038	-	i	-	-	-	⇔	-	-	-	-	-	044 - 032
049 - 040	-	-	-	-	-	⇔	-	v	-	-	-	044 - 036
049 - 042	-	-	-	-	-	⇔	-	v	-	-	-	044 - 038

-	Negative result
i	Inconclusive result
V	Variant result
+	Positive result

Source: GAO analysis of FBI repository data, 2014. | GAO-15-80

FBI officials stated that these results may have differed for a number of reasons, including uncertainty from the sampling process (sampling error) and uncertainty from the genetic test itself (stochastic error). Each step in the process the FBI used to collect, prepare, and test repository samples could have added uncertainty to the results of the genetic test.

As noted previously, before its searches, the FBI relied on laboratory officials to identify and select subsamples of distinct Ames strains for submission to the repository. The NAS report stated that the subpoena's instructions to laboratories for preparing samples were not precise enough to ensure consistent procedures for producing samples that would be most suitable for later comparisons. For example, the subpoena instructed laboratories to select a representative sample from each stock but did not provide guidance on how many cells or colonies to select. Although steps were taken in the genetic tests to standardize the number of cells being tested, the number of initial cells or colonies selected from each stock would have affected the probability of selecting material capable of producing the genetic markers.

This is particularly important because the mutations chosen as genetic markers were infrequent in the evidentiary material. For example, we interviewed the scientist who submitted the duplicate samples we identified above as having opposite results (all five negative versus all five positive). He told us that, in the presence of an FBI investigator, he had not followed the subpoena instructions when he selected the sample (047-002) that tested negative for all five genetic markers.

In addition to the selection methods we have discussed in this report, the methods used to prepare and test the repository samples could have introduced uncertainty to the results of the genetic tests. The NAS report stated that replication could have been used in the design of the FBI repository to provide measures of the uncertainty of the genetic tests.[62] Although laboratories were required to submit to the repository two samples from each stock, only one of those samples was tested for the genetic signature. Without replication, the FBI was unable to assess uncertainty in the results of the genetic tests in the context of testing actual repository samples.

[62] Replication is the selection of multiple samples from the same source. The NAS report suggested that laboratories should have been required to submit three or more samples of each stock to the repository so that repeated testing could be performed.

Because the FBI did not include measures of uncertainty when presenting the results of the genetic testing, questions have been raised about samples that tested positive for three or fewer genetic markers. For example, NAS stated that the FBI did not address false negative results and raised concern regarding the restriction of the statistical analyses to the repository samples that contained no inconclusive or variant results. NAS further highlighted 21 samples that contained an inconclusive or variant result and tested positive for 1, 2, or 3 genetic markers.

To illustrate the potential effect this uncertainty could have had on the interpretation of the results, we conducted an analysis using the estimates of false negative rates obtained from the additional replicate testing, combined with a sensitivity analysis accounting for the decision to restrict the statistical analyses to the 947 samples that contained no inconclusive or variant results. We computed a range of probabilities, given the observed results of the genetic testing, that each repository sample was selected from a stock that could have produced all four genetic markers.[63] We found an additional 16 repository samples with probabilities that exceeded a 1 percent chance of being selected from a stock that contained all four genetic markers. We determined that 15 of these 16 additional samples were selected from stocks held at the same two laboratories that were the source of one or more of the 8 samples that tested positive for all four genetic markers.

The remaining sample identified in our analysis was a sample that we had determined was independent from RMR-1029 and tested positive for the D marker. In addition, this sample was inconclusive for both the A1 and A3 markers and negative for the E marker. We computed a 0 to 19 percent range of probabilities for this sample, the maximum occurring when the model made the assumption that both inconclusive results for A1 and A3 markers were positive.

[63] The probability analysis made a number of important assumptions. We assumed that RMR-1029 and the evidentiary material contained all four variants at or above the LOD defined for each of the genetic tests, genetic variants in samples of the repository were at least as concentrated as in RMR-1029, the genetic tests were independent, and the prior distribution of the frequency of each genetic marker in the population was unknown. Additionally, since these replicate samples were selected in a controlled environment, false negative rates may have been under-estimated because they are not affected by variation in test results caused by the sampling procedures used to submit samples to the repository. These assumptions contribute to a conservative estimate of the probability of a source matching all four genetic markers. Appendix III describes the probability analysis in detail.

Additionally, the results of the genetic tests for this sample further highlight the importance of including measures of uncertainty. According to the transfer inventory records we reviewed and the laboratory official we interviewed, this sample was selected from a stock that was one of four copies of the same material. As shown in figure 6, the repository samples selected from the remaining three copies tested negative for all five genetic markers. This demonstrates that the genetic tests could have yielded different results for samples selected from the same material and, as the NAS stated, replication could have been used to provide measures of the uncertainty induced by these varying results.

Figure 6: The Results of Genetic Tests for the Four Samples from Copies of the Same Material

FBIR sample	A1	A3	D-2	D-1	E
066-015	i	i	+	+	-
066-013	-	-	-	-	-
066-034	-	-	-	-	-
066-038	-	-	-	-	-

-	Negative result
i	Inconclusive result
+	Positive result

Source: GAO analysis of FBI data, 2014. | GAO-15-80

The FBI Is Addressing the Need for Formal Statistical Expertise In Future Investigations

The FBI has taken steps to include statistical expertise in future investigations. The NAS report stated that the FBI appeared not to have sought formal statistical expertise early in this investigation and that similar investigations would benefit from including statistical expertise in their design and implementation. It noted that because many inferences depend on the design and analysis of complex data, the FBI should consult with expert statisticians throughout experimental design and planning, sample collection, sample analysis, and data interpretation. Further, the 2009 NRC report on strengthening forensic science in the United States highlighted the importance of statistical and quantitative proficiency for improving forensic science methods.

An FBI official told us that since the 2009 NRC report, the FBI has been building formal forensic statistical expertise both internally and externally.

For example, he said that the FBI laboratory division had created an internal statistical working group to examine the FBI's statistical needs in its forensic methods. The group included a professor of statistics visiting for 6 months to examine the statistical questions related to patterns, such as fingerprints, and also other science, such as chemistry and explosives. Additionally, the FBI has established a working relationship with members of the American Statistical Association's Ad-Hoc Advisory Committee on Forensic Science in order to discuss its statistical capacity. The FBI has also worked with other agencies to identify areas of statistical research needed for future investigations.

Scientific Gaps Remain Related to Verification, Validation, and Statistical Analyses and Research Is Ongoing

After the 2001 attack, the FBI did not conduct a lessons learned study but considers the NAS report to be one. The NAS report identified some scientific gaps related to the development of genetic tests and statistical analyses. In addition, we identified a key scientific gap that is related to the verification and validation of the genetic tests and the statistical analyses—that is, the significance of using genetic mutations in *B. anthracis* as genetic markers for analyzing evidentiary samples. DHS has funded some research on this gap but this research is not yet complete, and it is not yet known whether it will fully address the gap.

The NAS Report Identified Scientific and Technical Gaps

The FBI has not conducted a formal lessons-learned study of the scientific and technical methods it used in the investigation and thus has not specifically identified any scientific gaps in research related to the validation of genetic tests and statistical approaches. An FBI official stated that such a study was not needed because the 2001 incident was unique and the case is closed. This FBI official also told us that he considered the NAS report to be the lessons-learned study because it had identified several scientific gaps. For example, the NAS report indicated that the investigation lacked

1. a method for interpreting the genetic similarity between the attack spores in the letters and those in RMR 1029; and
2. an experimental design that included statistical input in the early stages of the investigation.

Nevertheless, the FBI does not necessarily agree with the scientific gaps that NAS highlighted in that report. However, the FBI stated in 2010 that the active dynamics of the microbial genome for any given species need to be understood—for example, the location on the genome of "hot spots" for mutation and diversity and whether there is a high rate of genetic mobility and change in any given species. Further, in September 2014, according to an FBI official, technology has changed since the investigation, and in the future genome sequencing will be used to analyze evidence samples.

A Key Gap Remains on the Significance of Using Genetic Mutations as Markers In Analyzing Evidentiary Samples

In addition to the gaps identified in the NAS report, we identified a key scientific gap that has not been fully addressed. This gap is related to the significance of using genetic mutations as genetic markers for analyzing evidentiary samples to determine their origins. Recognized by NAS, this issue is associated with the gaps it identified.[64] With respect to verification and validation, the genetic tests targeted specific DNA sequences of certain genetic mutations in their screening of the repository samples. The FBI used the results of the analysis of the repository screening by those tests to narrow the source of the attack spores. However, during the investigation, it was not known how stable genetic mutations were in a microbial genome or how significant they were as genetic markers.

We found that conditions causing the rise of the genetic mutations in the evidence were not known before or after validation or during the subsequent statistical analysis of the results of the repository screening. During the investigation, it was not known what conditions would have promoted or inhibited the presence of the genetic mutations at detectable levels. Such knowledge would have indicated whether they were associated with the evidence itself or with the culture practices normally used in a laboratory. Although FBI expert advisers recommended experiments, none were conducted at that time to attempt to obtain this information. Such experiments could have helped in understanding the evolution of these particular genetic mutations.

[64] According to NAS, the environmental effects (media, temperature, time) on the growth characteristics of bacteria and the likelihood of their developing morphs were not understood at that time and a method for interpreting the genetic similarity between the attack spores in the letters and those in RMR-1029 was lacking.

DHS-Funded Research on the Evolution of Morphological Variants is Ongoing

DHS has recognized the need for a methodology to determine how a material has been grown and produced and for obtaining information on the biology of agents, including their mutation rates and genome "hotspots" for mutation, so that their "relatedness" can be measured. In this context, an expert who reviewed this report stated that computational methods are also needed to reconstruct (or assemble) genome sequencing data so that the relationship between markers that are not independent, as is common in asexually reproducing bacterial genomes, can be inferred.

As a result, DHS has funded research that is intended to provide a better understanding of how morphological variants, or mutations, could emerge and evolve in bacterial genomes.[65] Some of the technologies involved in DHS's research, such as whole genome sequencing, are still evolving.[66] DHS-funded research includes studies of the population genetics of bacterial agents, including *B. anthracis*, at Northern Arizona University (NAU).[67] This research involves studies of diversity that include mutations among these agents. DHS's NBFAC is also studying genome sequencing methods. The purpose of these studies is to develop the capability to perform a metagenomic analysis of an entire sample using a hybrid-assembly. According to DHS, the field of "metagenomics," is broad but

[65] A bacterial genome is a bacterium's genetic information. It includes its complete set of genes.

[66] The Office of Science and Technology Policy (OSTP) published a national research strategy for microbial forensics in 2009 whose implementation it is coordinating with multiple federal agencies. However, according to OSTP, this strategy and the associated implementation plan are based not on scientific gaps related to the FBI's investigation of the 2001 anthrax attack but on advancing the broader issues of the field as a whole. The strategy is focused on the future—specifically, on efforts to improve sample collection, processing, preservation, recovery, and concentration of microbial pathogens and their signatures from collected samples for microbial forensic analyses. Interagency workgroups will be formed to address the areas the strategy identifies. Implementing the strategy has begun and assigns responsibility to specific federal agencies to conduct research set out in the strategy. See NSTC (National Science and Technology Council), *National Research and Development Strategy for Microbial Forensics*, Executive Office of the President, National Science and Technology Council (Washington, D.C.: Executive Office of the President, 2009). http://www.whitehouse.gov/files/documents/ostp/NSTC%20Reports/National%20MicroForensics%20R&DStrategy%202009%20UNLIMITED%20DISTRIBUTION.pdf (Accessed September 30, 2014).

[67] Bacterial population genetics is the study of the genetic diversity of bacterial populations. It attempts to define such diversity in terms of mutation, for example, and other factors.

unified by its focus on a community of genomes rather than individual isolates.[68] Such research is a step in the right direction, since the FBI has indicated that it is likely to use genome sequencing methods in future investigations to analyze evidence. However, since this research is ongoing it is not clear when it will close the gap or whether it can do so alone.

Conclusions

Although we identified several aspects of the FBI's scientific methods we reviewed that could be improved in a future investigation, we recognize that in 2001, the FBI was faced with an unprecedented case. Determining the source of the spores in the envelopes was complicated by many factors, including the uncertain provenance of samples in the FBI repository, an unknown mutation rate for *B. anthracis* under laboratory growth conditions, and the performance of the genetic tests under "real-world" conditions.

The genetic tests were generally verified, validated and demonstrated through the validation testing that they met the FBI's acceptance criteria, but the lack of a comprehensive approach—that is, a validation framework—allowed for differences in the contractors' approaches. Further, the results of the postvalidation testing raise questions about whether additional information could have been obtained during verification and validation and, thus, whether the validation testing could have been more rigorous. The use of a standardized approach to verification and validation from the beginning could have more definitively established the performance of all the genetic tests. It could have helped in communicating expectations clearly, ensuring confidence in results generated by any genetic tests developed.

DHS could be instrumental in developing a validation framework and future efforts using a framework could help achieve minimum performance standards during verification and validation, particularly under multiple contracts. Also, incorporating statistical analyses in the framework would allow the calculation of statistical confidence for interpreting the validation testing results.

[68] This method allows sampling of the genomes of microbes without culturing them. Before sequencing, the DNA is directly isolated from an environmental sample.

The FBI's statistical approach to its study design and plan, sample collection and analysis, and interpretation of data and scientific evidence lacked several important characteristics that could have strengthened the significance of that evidence. Although the complexity and novelty of the scientific methods at the time of the FBI's investigation made it challenging for the FBI to adequately address all these problems, the agency could have improved its approach by including formal statistical expertise early in the investigation and establishing a statistical framework that could identify and account for many of the problems. In future investigations, statistical expertise early in the investigation will help identify the importance and role of fully understanding the (1) evolution of the genetic markers, (2) sources of dependence between samples, and (3) uncertainty in the measurement tools used to identify a genetic signature. This expertise could influence an investigation's methods and strengthen the significance of scientific evidence.

A key scientific gap—how stable genetic mutations are in a microbial genome and thus their suitability as genetic markers—remains an issue. Lack of this knowledge has implications for both the development of genetic tests, or other investigative approaches and technologies, and the analysis of the results they generate. For example, how likely it is that the same genetic mutations will arise independently in separate cultures is currently unknown, and so is whether different culture conditions can change the ratio of the mutations significantly enough to provide a negative rather than a positive result. DHS-funded research into the evolutionary behavior of variants in the genome of *B. anthracis* and other microbial agents and the use of genome sequencing is a step in the right direction because the FBI is planning to use sequencing in future investigations to analyze all the material in evidence samples. However, in determining the significance of using mutations as genetic markers, an understanding is still needed about the stability of genetic mutations. DHS's ongoing research is likely to take several years and some of the technologies it entails, such as whole genome sequencing, are still evolving. Therefore, it is not clear when and whether this research alone will address this gap.

Recommendations

To ensure that a structured approach guides the validation of the FBI's future microbial forensic tests, we recommend that the Director of the Federal Bureau of Investigation work with the Secretary of Homeland Security to develop a verification and validation framework. The framework should be applied at the outset of an investigation involving an intentional release of *B. anthracis*, or any other microbial pathogen. It should (1) incorporate specific statistical analyses allowing the calculation of statistical confidence for interpreting the results and specifying the need for any additional testing to fully explore uncertainties relative to the type of genetic test being validated and (2) applied and adapted to a specific scenario and employs multiple contractors.

In addition, we recommend that the Director of the FBI establish a general statistical framework that would require input from statistical experts throughout design and planning, sample collection, sample processing, sample analysis, and data interpretation that can applied and adapted to address a specific scenario involving an intentional release of *B. anthracis* or any other microbial pathogen.

Agency Comments and Our Evaluation

We provided a draft of this report to the FBI and DHS for review and comment. The FBI provided written comments, which are reprinted in appendix IV. In its comments, the FBI agreed with our recommendations and stated that it had taken significant steps toward addressing them. In addition, the FBI provided technical comments that we have addressed in the body of our report as appropriate. DHS stated that it had no comments on the draft report.

With respect to the first recommendation, the FBI stated that “NBFAC programs have developed analytical capabilities in microbial forensics for numerous biological agents” in “support of investigations of the use or suspected use of biological weapons.” It stated that “these assays are validated and accredited under international standards (ISO17025) . . . .” According to the FBI, these capabilities, and those still being developed, “address part 2” of our recommendation “…applied and adapted to a specific scenario…” in as much as they represent capabilities addressing numerous biological agents and toxins. Further, the FBI stated that the NBFAC is pursuing the most current techniques of microbial genetic analyses and that some of these may soon be accredited.

The FBI added that it actively participates in the National Strategy for Countering Biological Threats, under which the agency has helped in “Establishing a National level research and development strategy and investment plan for advancing the field of microbial forensics.” Further, it stated that it is helping to maintain “the National Biological Forensics Analysis Center (NBFAC) as the Nation’s lead Federal facility for forensic analysis of biological material in support of law enforcement investigations,” which advances the field of microbial forensics through scientific workshops sponsored by the FBI. According to the FBI, such workshops have included work on interpreting microbial genetic data acquired by next generation sequencing platforms. The FBI stated that this work has included “statistical analyses of the confidence in base calling” using these platforms and “bioinformatic software.”   We recognize the importance of the FBI’s active participation in microbial forensic research and scientific workshops that address key issues related to the performance of emerging microbial forensic tests. We also recognize that establishing the error rates of genome sequencing platforms, which the FBI stated it may use in future investigations, would be an important step in verification and validation. Further, as we state in this report, developing a framework for verification and validation when employing multiple contractors in the same investigation could help standardize the process with minimum performance standards. Thus, we believe that the FBI’s continued work with DHS could help ensure the development of such a framework and improve its approaches to future investigations. A written plan could assist in the development of the framework.

With respect to the second recommendation, the FBI stated that scientists from the FBI and NBFAC participate in the Food and Drug Administration’s related efforts, the “Global Microbial Identifier” symposiums, “whose activities include statistical analyses for interpreting microbial genetic data in investigations of food-borne illness.” We recognize the importance of the FBI’s continued participation in research on the statistical interpretation of microbial genetic data. The evidence we present in this report suggests that if statistical expertise had been included early in the FBI’s investigation, it could have improved the significance of the collected microbial forensic evidence. By establishing a general statistical framework, the FBI will be able to provide some assurance that input from statistical experts will be included in future investigations so that they will benefit from statistical expertise. Developing such a framework could also be facilitated by a written plan. We believe that the actions that the FBI states it has taken are a step in the right direction toward addressing our two recommendations.

We are sending copies of the report to the FBI and DHS, appropriate congressional committees, and other interested parties. The report is also available at no charge on the GAO website at www.gao.gov.

If you or your staff have any questions about this report, please contact Timothy M. Persons, Ph.D. at (202) 512-6412 or personst@gao.gov. Contact points for our Office of Congressional Relations and Office of

Public Affairs appear on the last page of this report. Key contributors to the report are listed in appendix V.

T.M. Persons

Timothy M. Persons, Ph.D.
Chief Scientist

Appendix I: Objectives, Scope, and Methodology

The scope of our work was limited to a review of the scientific methods employed to validate the genetic tests used to screen the FBI's repository of Ames *B. anthracis* samples, the procedures used to identify and collect samples of Ames *B. anthracis* in the creation of the FBI's repository, and the statistical analyses and interpretation of the results of the genetic tests. We did not address any other scientific methods or any of the traditional investigative techniques used to support the FBI's conclusions in this case, and we take no position on the FBI's conclusions when it closed its investigation in 2010.

Our objective for this performance audit was to answer the following questions:

1. To what extent were the genetic assays used to screen the FBI repository of Ames samples scientifically verified and validated?
2. What are the characteristics of an adequate statistical approach for analyzing the repository samples and to what extent was the statistical approach used adequate? If not adequate, how could this approach be improved for future efforts?
3. What remaining scientific concerns and uncertainties, if any, regarding the validation of genetic assays and statistical approaches will need to be addressed in future analyses? What additional research, if any, would be helpful in resolving such scientific uncertainties in any future investigation?

To determine the extent to which the genetic tests were verified and validated, we collected and reviewed data regarding (1) the FBI's requirements for validation, (2) documentation from the FBI's contractors on their verification and validation testing, and (3) documentation from the FBI on the contractors' efforts to develop their genetic tests as well as results from the validation testing. We also reviewed related scientific literature and agency and industry standards and guidelines regarding the verification and validation of analytical methods, including real-time PCR-based tests for detecting *B. anthracis*, among others. We developed criteria for assessing the extent of the validation. We used references from agency standards, reports, and guidelines for validation and from scientific literature to identify the essential phases in an approach, or framework, for developing genetic tests. We compared what the FBI and its contractors had done to verify and validate the genetic tests against these phases and tasks.

Specifically, we reviewed the FBI's and its contractors' laboratory documentation to determine for each genetic test (1) the steps each took to verify the genetic tests' performance and conduct the FBI-administered validation, (2) whether the validation test results met the FBI's acceptance criteria and minimum requirements, and (3) whether the FBI's postvalidation testing of the genetic tests on the flask RMR-1029 provided further insights into the sensitivity and specificity of the genetic tests beyond those obtained by the validation. We also determined whether the processes the contractors' laboratories followed for verifying and validating their genetic tests were consistent. Finally, we reviewed the NAS report's observations on the performance of the genetic tests in screening the FBI's repository samples. We interviewed officials and scientists at the FBI contractors, the FBI, and elsewhere on how the genetic tests had been verified and validated, standards or guidelines had been applied, and the FBI's rationale for its requirements and acceptance of the five genetic tests as validated.

We also compared the validation test results with the results of the additional testing that was conducted after validation to determine if any additional information was provided on the performance characteristics of the genetic tests. We did not independently verify whether the contractors followed their quality assurance guidelines in developing, verifying, and validating their genetic tests, but we assumed that they did so from the documentation provided.

To determine the extent to which the statistical approach used for analyzing the repository samples was adequate, we used three approaches. First, we collected and analyzed documentation from the FBI, the three domestic laboratories searched by the FBI, and the contractor who did the statistical analyses. We reviewed contract records and conducted interviews with the FBI and laboratory officials. We conducted a literature review to collect relevant references from forensic science, statistics, epidemiology, and population genetics. Informed by the relevant literature, we identified and developed the set of characteristics that would be a statistical approach adequate to achieve the stated purposes of the FBI's statistical analyses.

We submitted the set of desirable characteristics described in this report to our experts and a subcommittee of the American Statistical Association's (ASA) Ad Hoc Advisory Committee on Forensic Science for their review and comment. To obtain information about how samples were selected from stocks and submitted to the repository, we reviewed the FBI subpoena protocols, conducted semi-structured interviews with officials, and collected relevant laboratory documentation from the three laboratories that the FBI searched.

Second, to obtain information about samples collected through the three follow-up searches, we interviewed FBI officials and reviewed the agency's documentation, conducted semi-structured interviews with officials from the three laboratories that the FBI searched, and reviewed relevant laboratory documentation. To identify duplicate samples in the repository, we compared the documentation of samples obtained through the searches to samples submitted through the subpoena process.

Third, to demonstrate the impact of the sensitivity of the genetic tests and data trimming assumptions made in the statistical analyses, we analyzed the FBI repository data and estimated false negative rates for each genetic test under repository conditions, using the post-validation results from replicate testing of RMR-1029 and evidentiary material. We conducted sensitivity analyses to examine the impact of data trimming assumptions made in the FBI's statistical analysis by varying the assumptions made to remove all inconclusive, no-growth, and variant results from the analysis. We computed conditional probabilities that a repository sample was selected from a stock containing all four morphs, given the observed combinations of genetic test results. We combined the probability analysis with the data trimming sensitivity analysis to compute a range of conditional probabilities for each repository sample. We identified the samples that had a maximum conditional probability of greater than 1 percent (nontrivial).

To assess the reliability of the FBI repository data, we summarized the data and compared the results to the contractor's final report on the statistical analysis and to published reports by the FBI and the National Academies to ensure external validity of the data. From the results of this testing, we found the data to be sufficiently reliable for the purposes of our review.

To determine any remaining scientific concerns and uncertainties regarding the validation of the genetic tests and statistical approaches that would need to be addressed in future analyses, we reviewed relevant federal agencies' and their contractors' documents, published literature, and industry documentation on the validation of polymerase chain-reaction based tests, such as those for detecting rare variants, and related scientific concerns and uncertainties that could affect a future investigation. We reviewed the contractors' final reports on the statistical analysis, reviewed contract documents, and interviewed FBI officials to identify where improvements to the approach could be made. In addition, we reviewed the Centers for Disease Control and Prevention's (CDC), the Animal and Plant Health Inspection Service's (APHIS), and the

Department of Defense's (DOD) select agent requirements for storing, handling, shipping, and maintaining inventory controls. We interviewed agency officials to determine if gaps exist in documenting important information about the provenance of *B. anthracis* stocks.

Further, to identify scientific concerns arising during the FBI's investigation of the validation of the genetic tests and statistical approaches, we reviewed pertinent documentation on scientific issues or problems the FBI and NAS had identified and their effect on the FBI's ability to validate the genetic tests or develop appropriate statistical approaches. Assisted by experts, we determined which gaps were significant and their potential effect on a future investigation with a similar scenario. We also interviewed officials and scientists at the contractors, the FBI, DHS, the National Bioforensic Analysis Center (NBFAC), DOD (at Dugway and USAMRIID), the Department of Energy's (DOE) Lawrence Livermore National Laboratory (LLNL), the Joint Genome Institute (JGI), EurekaGenomics, and the Executive Office of the President's Office of Science and Technology Policy (OSTP), regarding scientific challenges to genetic test validation, statistical analyses of the repository data, scientific gaps related to the FBI's investigation, and any federal research being conducted, or planned, to fill those gaps.

To determine additional research that would be helpful in resolving such scientific uncertainties in any future investigation, we reviewed documentation on research DHS is conducting to address any scientific gaps we found related to the validation of the genetic assays and issues related to the statistical analyses of the results of the repository screening. We reviewed the identified gaps and DHS's research and determined the progress that had been made to close them. Further, following on interviews with scientists and agency officials, and input by our experts, we determined whether any additional research is needed.

We asked scientists with expertise in public health and microbial forensic investigations to review and comment on a draft of our report. They included Jim Bristow, M.D., Deputy Director for Scientific Programs, DOE Joint Genome Institute; Karin S. Dorman, Associate Professor, Departments of Statistics and Genetics, Development, and Cell Biology, Iowa State University; George V. Ludwig, Ph.D., Deputy Principal Assistant for Research and Technology, U.S. Army Medical Research and Materiel Command; Jack Melling Ph.D., Director (retired), U.K. Centre for Applied Microbiology and Research, Porton Down, U.K.; Jeff Mohr, Ph.D., Chief (retired), Life Sciences Division, U.S. Army, Dugway

Proving Grounds; and Stephen Velsko, Ph.D., Senior Scientist and Associate Program Leader, Lawrence Livermore National Laboratory.

Finally, we asked a subcommittee of the American Statistical Association’s (ASA) Ad-hoc Advisory Committee on Forensic Science for its review and comment on the statistical aspects of a draft of our report. The subcommittee provided us with detailed comments that expressed general agreement with the statistical aspects of the draft, suggested changes to terminology related to the frequency with which microbial properties are present in a population, and suggested appropriate caveats and limitations to analyses we conducted. We incorporated these comments as appropriate throughout the report.

We conducted this performance audit from January 2013 to November 2014 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Appendix II: Performance Parameters Evaluated by Genetic Test

Parameter	A-1, A-3	D-1	D-2	E
1. Limit of detection	✓	✓	✓	✓
2. Sensitivity	✓	✓	✓	✓
3. Specificity and selectivity	✓	✓	✓	✓
4. Accuracy	✓	✓	✓	✓
5. Limit of quantitation	n.a.	n.a.	✓	✓
6. Linearity	n.a.	n.a.	✓	✓
7. Precision: repeatability	✓	✓	✓	o[a]
8. Intermediate precision	✓	✓	✓	o[a]
9. Precision: reproducibility	o	o	o	o
10. Range	✓	n.a.	✓	✓
11. Robustness	✓	✓	o	o[a]
12. Ruggedness	✓	✓	✓	o[a]
13. Other: Competition	o	✓	o	o[a]

Legend: ✓ = evaluated; o = not evaluated; n.a. = not applicable for qualitative tests.

Source: GAO analysis of contractor laboratory data. | GAO-15-80

[a] Contractor provided no information.

Appendix III: The Effect of Assay Sensitivity and Data Trimming Assumptions in the Statistical Analyses

To illustrate the potential effect of the sensitivity of the genetic tests and data trimming assumptions made in the statistical analyses, we analyzed the FBI repository data and estimated false negative rates for each assay under repository conditions using the results from postvalidation replicate testing of RMR-1029 and evidentiary material. We conducted a sensitivity analysis to examine the effect of data trimming assumptions in the FBI's statistical analysis by varying the assumptions to remove all inconclusive, no-growth, and variant results. We computed the conditional probabilities that a repository sample was selected from a stock containing all four genetic markers, given the observed combinations of results. We combined the probability analysis with the data trimming sensitivity analysis to compute a range of conditional probabilities for each repository sample. We then identified the samples that had a maximum conditional probability of greater than 1 percent (nontrivial).

To build a model to compute this probability, we defined the sample space of possible outcomes. There are 16 combinations for a binary measure of the presence (+) or absence (-) of each of the four genetic markers. Therefore we defined the possible outcomes for the four genetic markers (A1, A3, D, and E) as S1 through S16, as shown in figure 7.

Figure 7: Possible Outcomes for the Four Genetic Markers

S_1=(+ + + +)	S_5=(- + + +)	S_9=(+ - - +)	S_{13}=(- + - -)
S_2=(+ + + -)	S_6=(+ + - -)	S_{10}=(- + - +)	S_{14}=(- - + -)
S_3=(+ + - +)	S_7=(+ - + -)	S_{11}=(- - + +)	S_{15}=(- - - +)
S_4=(+ - + +)	S_8=(- + + -)	S_{12}=(+ - - -)	S_{16}=(- - - -)

Source: GAO. | GAO-15-80

The observed assay results have the 16 possible outcomes listed in figure 7. Since the goal of this analysis was to compute the probability that a repository sample had been selected from a stock that contained all four genetic markers, given the observed test result, we are interested in the probability of S_1, given the observed test result for a repository sample, $P(S1 \quad |obs)$. Using Bayes' theorem, this can be written as a posterior probability, $(S_1|obs) = \frac{P(obs|S_1)P(S_1)}{\sum_{i=1}^{16} P(obs|S_i)P(S_i)}$, where

$P(obs|S_1)$= “sample probability,” or the probability of an observed outcome given the stock contained all four genetic markers, and

$P(S_i)$= “prior probability,” or the probability of an outcome in the sample space.

Since there was no information on the distribution of the genetic markers in the population before the investigation, we assumed that the prior probabilities were equal, $P(S_i) = P\left(S_j\right)$, for i=1…16 and j$\neq i$, and the posterior probability simplified to $P(S_1|obs) = \frac{P(obs|S_1)}{\sum_{i=1}^{16} P(obs|S_i)}$. To compute the sample probability, $P(obs|S_1)$, we assumed that each test was independent and rewrote the probability as a linear combination of estimates of false negative rates for each assay (A1, A3, D, and E):

$$P(obs|S_1) = P_{A1}(obs_{A1}|+) * P_{A3}(obs_{A3}|+)P_D(obs_D|+) * P_E(obs_E|+)$$

We used statistics derived from the results of postvalidation replicate testing conducted on RMR-1029 and letter material to estimate false negative rates.

Figure 8 shows the breakdown of the results of the replicate testing.

Figure 8: Results of the Additional Replicate Testing

	RMR-1029						Letter mateiral				
Rep number	A1	A3	D-2	D-1	E	Rep number	A1	A3	D-2	D-1	E
1	+	+	+	+	+	1	+	+	+	+	-
2	+	+	+	+	+	2	+	+	+	+	-
3	+	+	+	+	+	3	+	+	+	+	-
4	+	+	+	+	+	4	+	+	+	+	-
5	+	+	+	+	+	5	+	+	+	+	-
6	+	+	+	+	+	6	+	+	+	+	-
7	+	+	+	-	+	7	+	+	+	+	+
8	+	+	+	+	+	8	+	+	+	+	+
9	+	+	+	+	+	9	+	+	+	+	+
10	i	+	i	i	+	10	+	+	+	+	+
11	+	+	+	+	+	11	+	+	+	+	+
12	i	+	-	i	+	12	+	+	+	+	+
13	i	+	i	-	+	13	+	+	+	+	+
14	-	+	+	+	+	14	+	+	+	+	+
15	-	+	+	+	+	15	+	+	+	+	+
16	+	+	+	+	+	16	+	+	+	+	-
17	+	+	+	+	+	17	+	+	+	+	+
18	-	+	+	+	+	18	+	+	+	+	+
19	+	+	+	+	+	19	+	+	+	+	+
20	-	i	-	-	+	20	+	+	+	+	+
21	+	+	+	+	+	21	+	+	+	+	+
22	-	+	+	+	+	22	+	+	+	+	+
23	+	+	+	+	+	23	+	+	+	+	+
24	-	+	-	-	+	24	+	+	+	+	+
25	i	+	+	+	+	25	+	+	+	+	+
26	+	+	+	+	+	26	+	+	+	+	+
27	-	+	+	+	+	27	+	+	+	+	+
28	+	+	+	+	+	28	+	+	+	+	+
29	-	+	+	+	+	29	+	+	+	+	+
30	-	+	-	i	+	30	+	+	+	+	+

- Negative result
i Inconclusive result
+ Positive result

Source: GAO analysis of FBI data, 2014. | GAO-15-80

The sensitivity analysis examined the effect of two data trimming decisions made in the FBI's statistical analysis of the repository samples—the choice of D assay results and the treatment of inconclusive results. The D marker was typed by two assay procedures (D-1 and D-2), only one of which (D-2) the FBI used in its analysis. The Statistical Analysis Report was restricted to the analysis of 947 samples that contained no inconclusive or variant results and, therefore, excluded 112 samples. To explore the potential effect of the inconclusive exclusion on the probabilities of observing all four morphs, we explored three possible outcomes for inconclusive results. We treated all inconclusive results as first positive and then negative, and then we excluded the inconclusive results from the analysis. The sensitivity analysis examined the six different combinations of outcomes, the two D assay possibilities, and the three potential outcomes of the inconclusive data.

The computation included all 1,059 repository samples and varied the assumptions made around data trimming from most to least conservative. The results for each set of estimated false negative rates show that 7 of the 16 possible outcomes of the genetic testing had a range of probabilities that included values exceeding a 1 percent chance of being selected from a stock that contained all four genetic markers (table 8).

Table 8: Probability Ranges for 16 Possible Outcomes from Two Different False Negative Estimates

16 possible outcomes	1: RMR-1029 material		2: Letter material	
	Min	Max	Min	Max
(+ + + +)	1.000	1.000	1.000	1.000
(+ + + -)	0.000	0.000	0.189	0.189
(+ + - +)	0.040	0.189	0.000	0.000
(+ + - -)	0.000	0.000	0.000	0.000
(+ - + +)	0.000	0.032	0.000	0.000
(+ - + -)	0.000	0.000	0.000	0.000
(+ - - +)	0.000	0.006	0.000	0.000
(+ - - -)	0.000	0.000	0.000	0.000
(- + + +)	0.226	0.302	0.000	0.000
(- + + -)	0.000	0.000	0.000	0.000
(- + - +)	0.009	0.057	0.000	0.000
(- + - -)	0.000	0.000	0.000	0.000
(- - + +)	0.000	0.010	0.000	0.000
(- - + -)	0.000	0.000	0.000	0.000

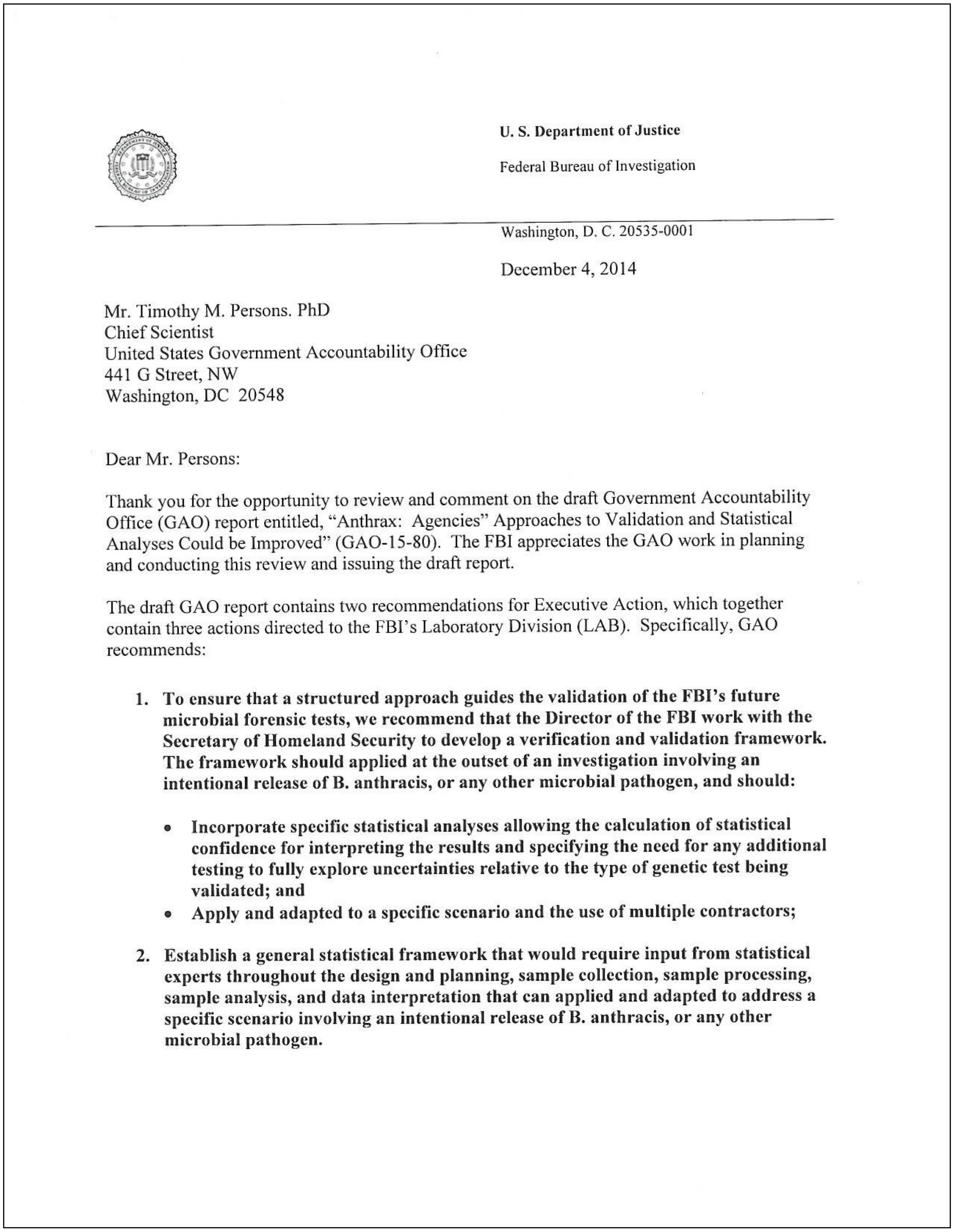
16 possible outcomes	1: RMR-1029 material		2: Letter material	
	Min	Max	Min	Max
(- - - +)	0.000	0.002	0.000	0.000
(- - - -)	0.000	0.000	0.000	0.000

Source: GAO. | GAO-15-80

Further, when we computed the probabilities for the repository samples, we found that only a small subset of the 1,059 repository samples had a range of probabilities that included values that exceeded a 1 percent chance of being selected from a stock that contained all four genetic markers. Specifically, we identified 24 repository samples, including the 8 that tested positive for all four genetic markers, which had a nontrivial chance of being selected from a stock that contained all four genetic markers.

By using estimates of false negative rates from the results of the postvalidation replicate tests on RMR-1029 and the letter material, we have made an assumption that the genetic variants in all samples in the FBI repository were at least as concentrated as in RMR-1029 or the letter material. Additionally, since these replicate samples were selected in a controlled environment, false negative rates may have been underestimated because they are not affected by variation in test results caused by the sampling procedures used to submit samples to the repository. These assumptions contribute to a conservative estimate of the probability of a source matching all four genetic markers.

Appendix IV: Comments from the FBI



U. S. Department of Justice

Federal Bureau of Investigation

Washington, D. C. 20535-0001

December 4, 2014

Mr. Timothy M. Persons. PhD
Chief Scientist
United States Government Accountability Office
441 G Street, NW
Washington, DC 20548

Dear Mr. Persons:

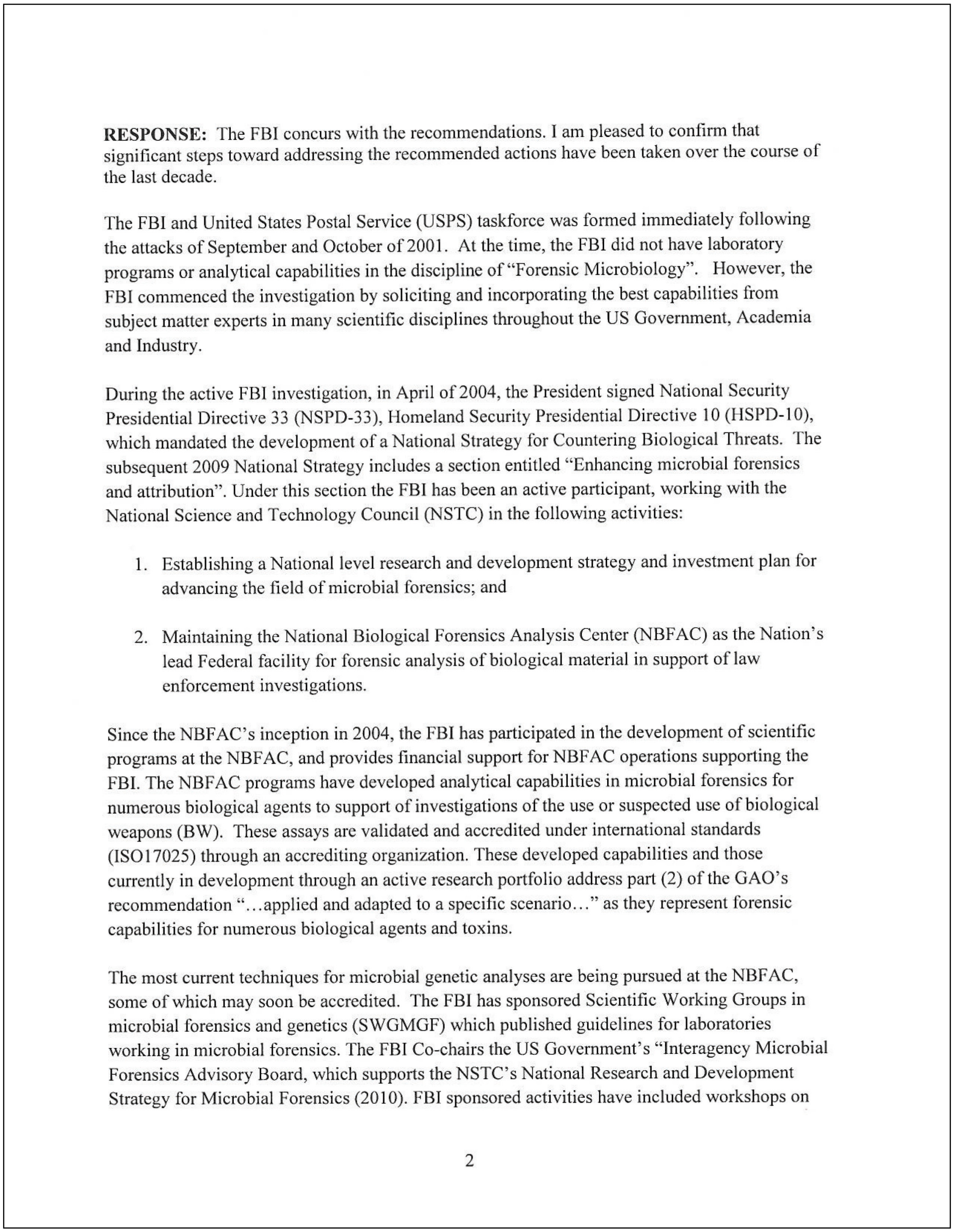
Thank you for the opportunity to review and comment on the draft Government Accountability Office (GAO) report entitled, "Anthrax: Agencies" Approaches to Validation and Statistical Analyses Could be Improved" (GAO-15-80). The FBI appreciates the GAO work in planning and conducting this review and issuing the draft report.

The draft GAO report contains two recommendations for Executive Action, which together contain three actions directed to the FBI's Laboratory Division (LAB). Specifically, GAO recommends:

1. **To ensure that a structured approach guides the validation of the FBI's future microbial forensic tests, we recommend that the Director of the FBI work with the Secretary of Homeland Security to develop a verification and validation framework. The framework should applied at the outset of an investigation involving an intentional release of B. anthracis, or any other microbial pathogen, and should:**

    - **Incorporate specific statistical analyses allowing the calculation of statistical confidence for interpreting the results and specifying the need for any additional testing to fully explore uncertainties relative to the type of genetic test being validated; and**
    - **Apply and adapted to a specific scenario and the use of multiple contractors;**

2. **Establish a general statistical framework that would require input from statistical experts throughout the design and planning, sample collection, sample processing, sample analysis, and data interpretation that can applied and adapted to address a specific scenario involving an intentional release of B. anthracis, or any other microbial pathogen.**



RESPONSE: The FBI concurs with the recommendations. I am pleased to confirm that significant steps toward addressing the recommended actions have been taken over the course of the last decade.

The FBI and United States Postal Service (USPS) taskforce was formed immediately following the attacks of September and October of 2001. At the time, the FBI did not have laboratory programs or analytical capabilities in the discipline of "Forensic Microbiology". However, the FBI commenced the investigation by soliciting and incorporating the best capabilities from subject matter experts in many scientific disciplines throughout the US Government, Academia and Industry.

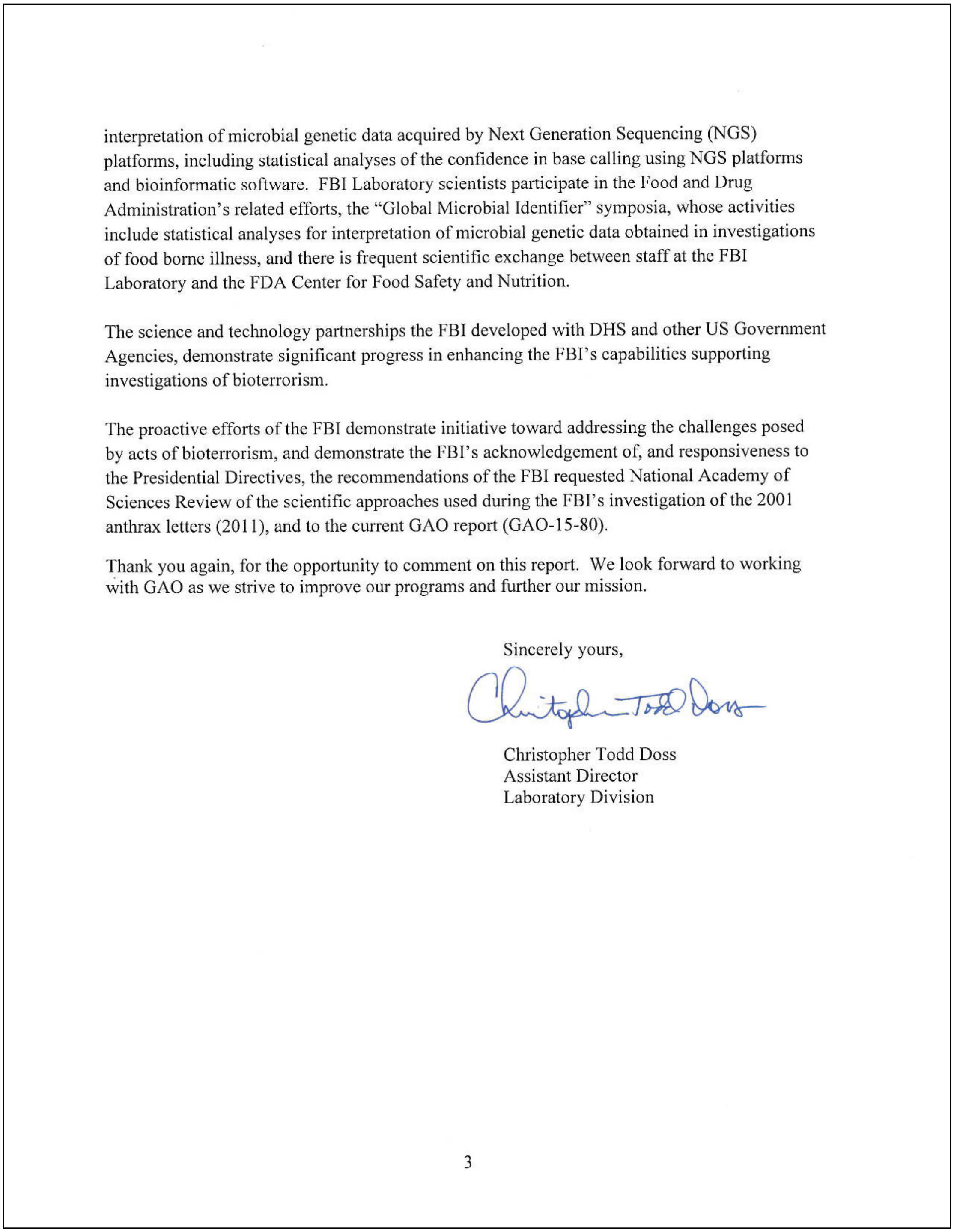
During the active FBI investigation, in April of 2004, the President signed National Security Presidential Directive 33 (NSPD-33), Homeland Security Presidential Directive 10 (HSPD-10), which mandated the development of a National Strategy for Countering Biological Threats. The subsequent 2009 National Strategy includes a section entitled "Enhancing microbial forensics and attribution". Under this section the FBI has been an active participant, working with the National Science and Technology Council (NSTC) in the following activities:

1. Establishing a National level research and development strategy and investment plan for advancing the field of microbial forensics; and

2. Maintaining the National Biological Forensics Analysis Center (NBFAC) as the Nation's lead Federal facility for forensic analysis of biological material in support of law enforcement investigations.

Since the NBFAC's inception in 2004, the FBI has participated in the development of scientific programs at the NBFAC, and provides financial support for NBFAC operations supporting the FBI. The NBFAC programs have developed analytical capabilities in microbial forensics for numerous biological agents to support of investigations of the use or suspected use of biological weapons (BW). These assays are validated and accredited under international standards (ISO17025) through an accrediting organization. These developed capabilities and those currently in development through an active research portfolio address part (2) of the GAO's recommendation "…applied and adapted to a specific scenario…" as they represent forensic capabilities for numerous biological agents and toxins.

The most current techniques for microbial genetic analyses are being pursued at the NBFAC, some of which may soon be accredited. The FBI has sponsored Scientific Working Groups in microbial forensics and genetics (SWGMGF) which published guidelines for laboratories working in microbial forensics. The FBI Co-chairs the US Government's "Interagency Microbial Forensics Advisory Board, which supports the NSTC's National Research and Development Strategy for Microbial Forensics (2010). FBI sponsored activities have included workshops on



interpretation of microbial genetic data acquired by Next Generation Sequencing (NGS) platforms, including statistical analyses of the confidence in base calling using NGS platforms and bioinformatic software. FBI Laboratory scientists participate in the Food and Drug Administration's related efforts, the "Global Microbial Identifier" symposia, whose activities include statistical analyses for interpretation of microbial genetic data obtained in investigations of food borne illness, and there is frequent scientific exchange between staff at the FBI Laboratory and the FDA Center for Food Safety and Nutrition.

The science and technology partnerships the FBI developed with DHS and other US Government Agencies, demonstrate significant progress in enhancing the FBI's capabilities supporting investigations of bioterrorism.

The proactive efforts of the FBI demonstrate initiative toward addressing the challenges posed by acts of bioterrorism, and demonstrate the FBI's acknowledgement of, and responsiveness to the Presidential Directives, the recommendations of the FBI requested National Academy of Sciences Review of the scientific approaches used during the FBI's investigation of the 2001 anthrax letters (2011), and to the current GAO report (GAO-15-80).

Thank you again, for the opportunity to comment on this report. We look forward to working with GAO as we strive to improve our programs and further our mission.

Sincerely yours,

Christopher Todd Doss

Christopher Todd Doss
Assistant Director
Laboratory Division

Appendix V: GAO Contact and Staff Acknowledgments

GAO Contact

Timothy M. Persons (Chief Scientist), (202) 512-6412 or personst@gao.gov.

Staff Acknowledgments

In addition to the contact named above, Sushil Sharma, Assistant Director, Pille Anvelt, James Ashley, Hazel Bailey, Amy Bowser, Mae Liles, Jan Montgomery, Penny Pickett, and Elaine Vaurio also made key contributions to this report.

Related GAO Products

Anthrax: DHS Faces Challenges in Validating Methods for Sample Collection and Analysis GAO-12-488 (Washington D.C.: July 31, 2012).

Federal Agencies Have Taken Some Steps to Validate Sampling Methods and to Develop a Next-Generation Anthrax Vaccine, GAO-06-756T (Washington D.C.: May 9, 2006).

Anthrax Detection: Agencies Need to Validate Sampling Activities in Order to Increase Confidence in Negative Results, GAO-05-251 (Washington D.C.: March 31, 2005).

U.S. Postal Service: Better Guidance Is Needed to Ensure an Appropriate Response to Anthrax Contamination, GAO-04-239 (Washington D.C.: September 9, 2004).

U.S. Postal Service: Issues Associated with Anthrax Testing at the Wallingford Facility, GAO-03-787T (Washington D.C.: May 19, 2003).

U.S. Postal Service: Better Guidance Is Needed to Improve Communication Should Anthrax Contamination Occur in the Future, GAO-03-316 (Washington D.C.: April 7, 2003).

(460615)