Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation

David A. Rasko[a], Patricia L. Worsham[b], Terry G. Abshire[b], Scott T. Stanley[c,1], Jason D. Bannan[d], Mark R. Wilson[d,2], Richard J. Langham[c], R. Scott Decker[c,3], Lingxia Jiang[a,4], Timothy D. Read[e], Adam M. Phillippy[f], Steven L. Salzberg[f], Mihai Pop[f], Matthew N. Van Ert[g,h], Leo J. Kenefic[g,h,5], Paul S. Keim[g,h], Claire M. Fraser-Liggett[i], and Jacques Ravel[a,6]

[a]Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201; [b]US Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD 21702-5011; [c]Washington Field Office, Federal Bureau of Investigation, Washington, DC 20535; [d]Federal Bureau of Investigation Laboratory, Quantico, VA 22134-5163; [e]Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA 30322; [f]Department of Computer Sciences, University of Maryland, College Park, MD 20742; [g]Northern Arizona University, Flagstaff, AZ 86011; [h]Translational Genomics Research Institute, Flagstaff, AZ 85004; and [i]Institute for Genome Sciences, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201

Edited* by Rita R. Colwell, University of Maryland, College Park, MD, and approved February 2, 2011 (received for review November 12, 2010)

Before the anthrax letter attacks of 2001, the developing field of microbial forensics relied on microbial genotyping schemes based on a small portion of a genome sequence. Amerithrax, the investigation into the anthrax letter attacks, applied high-resolution whole-genome sequencing and comparative genomics to identify key genetic features of the letters' *Bacillus anthracis* Ames strain. During systematic microbiological analysis of the spore material from the letters, we identified a number of morphological variants based on phenotypic characteristics and the ability to sporulate. The genomes of these morphological variants were sequenced and compared with that of the *B. anthracis* Ames ancestor, the progenitor of all *B. anthracis* Ames strains. Through comparative genomics, we identified four distinct loci with verifiable genetic mutations. Three of the four mutations could be directly linked to sporulation pathways in *B. anthracis* and more specifically to the regulation of the phosphorylation state of Spo0F, a key regulatory protein in the initiation of the sporulation cascade, thus linking phenotype to genotype. None of these variant genotypes were identified in single-colony environmental *B. anthracis* Ames isolates associated with the investigation. These genotypes were identified only in *B. anthracis* morphotypes isolated from the letters, indicating that the variants were not prevalent in the environment, not even the environments associated with the investigation. This study demonstrates the forensic value of systematic microbiological analysis combined with whole-genome sequencing and comparative genomics.

The causative agent of anthrax in mammals, *Bacillus anthracis*, is a recently emerged pathogen with characteristics of a young evolutionary group (1). It occurs primarily as quiescent spores that can persist for long periods of time in the environment between rapid proliferation growth phases within a host (2). Because of this episodic and highly specialized life cycle, the species is characterized by relatively little genetic variation (3–5). The most divergent *B. anthracis* strains are believed to share >99.99% nucleotide sequence identity (1, 6). *B. anthracis* has long been considered a potential biological weapon because of the stability of its spores, its potential for aerosolization, and its high rate of infectivity and lethality. The anthrax letter attacks in the fall of 2001 affirmed the potential of *B. anthracis* as a biological weapon (7, 8).

In the investigation that followed, scientists and investigators were confronted for the first time with the challenge of attributing the spore preparation of a genetically homogeneous species (used in the letters) to a potential source (4, 9). Genotyping methods, such as multiple-locus variable-number tandem repeat analysis and an additional typing system based on 12 canonical SNPs, could identified colony isolates as the Ames strain yet were insufficient to distinguish between isolates (1, 3, 4, 10–12). Because of the limitations of existing typing systems, it was thought that only whole-genome sequencing of Ames isolates would provide the level of discrimination required for microbial forensics and attribution. In recent years, microbial whole-genome sequencing has evolved from examining "model organisms" to becoming an important tool for the comprehensive comparative analysis of pathogens, with applications to the fields of bacterial pathogenesis and vaccine development (13), epidemiology (14), and, now, microbial forensics.

In 2003, The Institute for Genomic Research completed the whole chromosome sequence and analysis of the plasmidless *B. anthracis* Ames Porton (15), because this strain had been cured of both virulence plasmids, pXO1 and pXO2, by heat (43 °C) and novobiocin treatment, respectively (16). Comparison of the genome sequence of Ames Porton to the draft genome sequence of the *B. anthracis* strain isolated from the first victim of the 2001 mail anthrax attacks in Florida (Ames Florida strain A2012) revealed a limited number of novel and potentially diagnostic genetic variations that included chromosomal SNPs and insertions/deletions (INDELs) (9). Genetic variation was also identified in the plasmid sequences of Ames Florida by comparison with pXO1 and pXO2 from *B. anthracis* Sterne (pXO1+, pXO2−) and Pasteur (pXO1−, pXO2+), respectively, the only two representative *B. anthracis* plasmid sequences (17, 18). It became clear that the chromosomal and plasmid genetic variations were not characteristic of Ames Florida but rather were specific to Ames Porton because of the mutagenic effects of the plasmid curing process. This study highlighted the need for a high-quality genome sequence of a fully virulent *B. anthracis* Ames.

B. anthracis Ames was isolated in Sarita, TX, from a dead 14-mo-old Beefmaster heifer. It was acquired as a tryptose agar slant in 1981 by researchers at the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) from the Texas A&M Veterinary Medicine and Diagnostic Laboratory Bacteri-

Author contributions: D.A.R., P.L.W., T.G.A., S.T.S., M.R.W., R.J.L., R.S.D., T.D.R., C.M.F.-L., and J.R. designed research; D.A.R., P.L.W., T.G.A., S.T.S., L.J., and J.R. performed research; D.A.R., P.L.W., T.G.A., T.D.R., A.M.P., S.L.S., M.P., P.S.K., C.M.F.-L., and J.R. contributed new reagents/analytic tools; D.A.R., J.D.B., M.R.W., R.J.L., R.S.D., A.M.P., S.L.S., M.P., M.N.V.E., L.J.K., P.S.K., and J.R. analyzed data; and D.A.R., S.T.S., J.D.B., M.R.W., R.S.D., A.M.P., S.L.S., M.P., P.S.K., and J.R. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]Present address: Federal Bureau of Investigation, Critical Incident Response Group, Quantico, VA 22135.

[2]Present address: Department of Chemistry and Physics, Forensic Science Program, Western Carolina University, Cullowhee, NC 28723.

[3]Present address: Headquarters, Federal Bureau of Investigation, Washington, DC 20535.

[4]Present address: Canon Life Sciences, Inc., Rockville, MD 20850.

[5]Present address: University of Maryland School of Medicine, Baltimore, MD 21201.

[6]To whom correspondence should be addressed. E-mail: jravel@som.umaryland.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1016657108/-/DCSupplemental.

ology Department. At USAMRIID, it was selected as a challenge strain for anthrax vaccine preparations. This particular accession of *B. anthracis* Ames is believed to be the earliest known and progenitor of *all* Ames samples used as research tools in laboratories around the world; hence, it is an important, fully virulent reference for the Ames genotype (19). This material is hereafter referred to as *B. anthracis* Ames Ancestor. The complete high-quality genome sequence of *B. anthracis* Ames Ancestor (9) is the reference used in all comparative analyses in this study.

In September and October 2001, at least four envelopes containing *B. anthracis* spores were mailed through the U.S. Postal Services. The Federal Bureau of Investigation (FBI) recovered two letters postmarked September 18, 2001, addressed to the *New York Post* (NY Post) and to Tom Brokaw at NBC Television in New York City. Two additional letters recovered by the FBI were postmarked October 9, 2001, and sent to the Washington, DC, offices of Senators Daschle and Leahy (6–8). It is presumed that an additional letter was mailed to the American Media International (AMI) offices in Boca Raton, FL; however, no letter was recovered (7, 8). At least 22 victims contracted anthrax as a result of the mailings: 11 individuals contracted inhalation anthrax, with 5 of these infections resulting in fatalities; another 11 individuals suffered cutaneous anthrax. In addition, 31 persons tested positive for exposure to *B. anthracis* spores. All of the exposure and infection cases were attributed to the mailings based on a combination of timing, location, and place of employment as well as ultimately by genotyping that matched the *B. anthracis* Ames strain (7, 8). The resulting scientific investigation applied principles of microbial forensics (20) to fully characterize the threat letters and their content by using numerous physical, chemical, and genetic analyses. This study demonstrates that detailed microbiological observation and genomic analyses, combined with the principles of microbial population genetics and microbial forensics, were instrumental to the development of assays that established a link between the *B. anthracis* spores recovered from the letters and potential biological sources.

Results

Microbiological Identification. By using microbiological assays outlined in the *SI Materials and Methods*, evidentiary spore samples from the letters sent to the NY Post and Senators Leahy and Daschle were demonstrated to contain a subpopulation of minor variants that were morphologically distinct from the predominant wild-type colonies of *B. anthracis* Ames Ancestor (Table 1 and Fig. 1). Insufficient spore material was recovered from the letter addressed to Tom Brokaw at NBC; thus, it was not included in this analysis. Examination of the colonies formed on sheep blood agar (SBA) resulted in the identification of four distinct morphological variants (morphotypes) that we designated A, B, C/D, and E from each of the material analyzed. All morphotypes were sensitive to γ-phage. The colonies of each of the variants were colored yellow or yellow-gray rather than the typical gray-white of a wild-type colony of *B. anthracis*. Morphologically, three of the four variant colonies (A, B, and C/D) were large and spreading compared with wild-type colonies, whereas morphotype E exhibited round, compact, and dense colonies (Fig. 1). For each of the morphotypes, aberrant color and morphology were maintained on SBA medium upon subculturing. Secondary cultures of each morphological variant were assayed after 24 h of growth for sporulation on two media, new sporulation medium (NSM) and Leighton-Doi (L-D) medium, at two temperatures. Congo red binding, previously demonstrated to be associated with the ability to sporulate, was also evaluated (21). The ability to sporulate was consistently reduced in each of the variants compared with *B. anthracis* Ames Ancestor (Table 1). These phenotypic differences were maintained and stable in subcultures of the variants. These oligosporogenic morphotypes were consistently isolated from evidentiary spore samples from each of the letters analyzed, although they were never observed in spore preparations of *B. anthracis* Ames Ancestor evaluated in parallel with identical methods, demonstrating that cultures of *B. anthracis* Ames do not readily generate these specific mutations to detectable levels under assay conditions. Wild type and colonies with variant phenotypes recovered from each of the evidentiary spore samples were selected for further genetic analysis, including whole-genome sequencing, and kept for archival purposes.

Sequence Analysis of Morphological Variants. To identify the underlying genetic differences responsible for the variant phenotypes, whole-genome sequencing was performed. The wild type and morphotype B from the NY Post letter (PL) evidentiary spore sample (identified as PL1 and PL9, respectively) as well as morphotype A isolated from the material recovered from the Leahy letter (LL10) were sequenced to completion (every base at high quality and coverage), whereas the other isolates were sequenced to high-quality draft level, 12× sequence coverage (Table S1). Each of these draft genomes sequences were assembled in 5–84 contigs and 1–38 scaffolds (Table S2). Because the underlying genetic variations associated with each morphotype were consistent in the draft and closed genome sequences of variants from the Leahy and NY Post letters, no variants isolated from the Daschle letter (DL) were sequenced. Whole-genome shotgun sequencing was combined with sophisticated comparative bioinformatics tools that incorporated statistical validation

Table 1. Morphotype description

Phenotype examined	Wild type	A	B	C/D	E (Opaque)
Color	Gray-white	Yellow-gray	Yellow	Gray-yellow	Yellow
Colony morphology	Medusa, compact	Spreading, very flat, uniform	Spreading	Somewhat spreading	Round, compact, dense
Congo red binding*	3+	1+	1+	2+	ND
Pick plates	Compact, pitted	Largest, slightly hemolytic, moist "bull's-eye" center	Large, hemolysis around and under colony, very uniform	Concentric rings, slightly hemolytic, larger than wild type	Very compact, round, not pitted
NSM morphology	Shiny, pitted	Very thin growth, nonpitted	Shiny, nonpitted	Shiny, slightly pitted	Shiny, slightly pitted
NSM pigmentation	1+	4+	3+	2+	2+
Sporulation L-D (37 °C)	4+	±	±	1–2+	±
Sporulation SBA (37 °C)	4+	1+	1+	2–3+	±
Temperature-sensitive sporulation†	No	No	No	No	Yes

ND, not determined.
*For Congo red testing, see Worsham and Sowers (21).
†Sporulation was rated as 2–3+ at 28 °C.

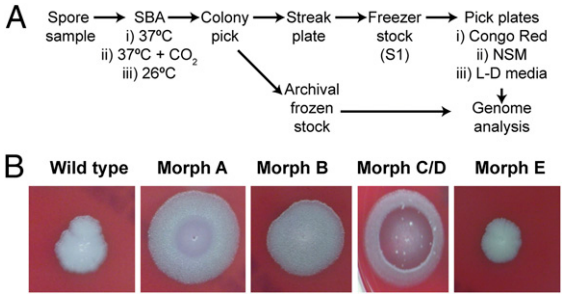

Fig. 1. Microbiological identification of morphological variants of *B. anthracis* Ames. (*A*) Flowchart used to process the evidentiary material for the identification of the morphological variants. In all cases, the morphotypes are altered in the sporulation phenotype and colony morphology. (*B*) Image of a representative colony of each of the morphotypes grown on SBA.

of an identified sequence variation. We used two criteria to validate the identified sequence variants: (*i*) sequence read coverage (number of sequence reads covering the variable region) of at least 3× and (*ii*) sequence quality [Phred score (22, 23)] of each individual read level base pair used to call the consensus sequence of >30. By using this strategy, high-quality sequence variants were discovered in the draft and closed genomes of the *B. anthracis* Ames morphotypes isolated from the letters' spore powders (Table S1).

Wild-Type *B. anthracis* Ames Isolates. No genetic variants were found when the closed genome sequence of the wild-type isolate from the Leahy letter was compared with that of *B. anthracis* Ames Florida (9) and Ames Ancestor (19). The 5,237,519 bp of the chromosome and the two plasmids pXO1 and pXO2 were identical. Similarly, the draft genome sequence of the wild-type isolate from the NY Post letter was identical to *B. anthracis* Ames Ancestor. In contrast, each of the morphological variants analyzed contained polymorphic loci.

Morphotype A. Morphotype A was identified in all three available evidentiary spore samples. These colonies were large, oligosporogenic, and pigmented (Table 1 and Fig. 1). The comparison of the closed genome of a Leahy letter morphotype A isolate (LL10) to that of *B. anthracis* Ames Ancestor revealed no genetic differences. This finding prompted further examination of the genome assembly that identified a region in LL10 where several sequence read pairs were in violation of their expected insert size and clone orientation as well as unassembled reads with a unique sequence junction (Fig. 2*A*). This observation is characteristic of a misassembled tandem repeat sequence, where the tandem repeat sequence is incorrectly collapsed by the assembler into a single repeat unit (Fig. 2*A*). The genetic variation identified in the LL10 genome sequence is a 2,023-bp repeat, which was located upstream of the fourth copy of the rRNA gene operon (*rrsD*, *rrlD*, *rrfD*) and included GBAA_0150 and GBAA_0151, encoding a putative polysaccharide deacetylase and a hypothetical protein, respectively (Fig. 2*B*). Analysis of the draft genome sequence of the NY Post letter morphotype A (PL10) identified a similar misassembly at the same locus; however, the repeat unit was larger (2,607 bp) and did not have the same boundaries as those in LL10. Interestingly, both repeated sequence units contained the hypothetical protein GBAA_0151.

A PCR-based assay was designed to confirm the bioinformatics findings. PCR primers 1572F and 1572R, both internal to GBA_0151 (Fig. 2*B*), were designed to produce an amplicon only if a duplication was present because the two primers point away from each other in a single instance of GBAA_0151. In this assay, the size of the observed amplicon was the size of the repeated sequence unit. This PCR assay was applied to genomic DNA isolated from *B. anthracis* Ames Ancestor, *B. anthracis* Ames Porton, and morphotype A isolates from the Leahy (LL10), NY Post (PL10), and Daschle (DL10) letters. As expected, *B. anthracis* Ames Ancestor did not contain the duplication at this locus, whereas an amplicon was observed with each of the morphotype A isolates tested, indicating the presence of the duplication. Interestingly, the size of the duplicated regions was different in each of the letter variants tested (Fig. 2*C*). Sequencing of each of the PCR products identified the exact location of the duplication junctions and confirmed the sequence assembly data. The Daschle letter morphotype A contained an 822-bp repeated sequence unit. Surprisingly, *B. anthracis* Ames Porton was positive in this assay and contained a 1,260-bp duplication at this locus. This information led us to reexamine the genome sequence assembly of *B. anthracis* Ames Porton, where misassembled reads were found and the assembly was corrected (GenBank accession no. AE016879).

Further analysis of the draft genome sequence of NY Post letter morphotype A (PL10) revealed an additional SNP at position 3,837,162 (all nucleotide locations are in reference to the *B. anthracis* Ames Ancestor genome sequence). This nonsynonymous point mutation (AGA→ATA; R→I) altered the final codon of gene GBAA_4191, encoding a predicted TrkA family potassium-uptake protein. The SNP position was covered by four agreeing reads and was of high quality (median quality score of 32.25). A PCR/sequencing assay (Table S3 and S4) was designed that amplified a 571-bp amplicon surrounding the SNP position. The assay revealed the unique association of the SNP with the NY Post morphotype A isolate and was not found in any other isolates tested.

Morphotype B. The genome of a NY Post letter morphotype B isolate (PL9) was sequenced to completion and compared with that of *B. anthracis* Ames Ancestor. The comparative genome sequence analysis revealed a single high-quality, high-coverage SNP (T→C) located at position 5,065,092. This SNP was located in the intergenic region between divergently transcribed genes GBAA_5581, annotated as stage 0 sporulation protein F (*spo0F*), and GBAA_5582, a conserved hypothetical protein (Fig. 3 *Lower Left*). Comparison of the draft genome sequence of Leahy letter morphotype B (LL9) revealed an identical SNP. In the NY Post letter morphotype B, the SNP position was covered by five agreeing sequence reads (median quality of 41.5), whereas the same SNP position was covered by nine high-quality sequence reads (median quality of 43) in the Leahy letter morphotype B genome sequence (Fig. S1). By using a 466-bp PCR/sequencing-based assay (Table S3 and S4), the presence of this SNP specific to morphotype B was reconfirmed in the NY Post and Leahy letter morphotype B isolates and an identical SNP in the Daschle letter morphotype B isolate (DL9). Neither *B. anthracis* Ames Ancestor nor Ames Porton carried this SNP, suggesting that this genetic variant was specific to morphotype B.

Morphotypes C/D. The genome of a Leahy letter morphotype C isolate (LL6) was sequenced to high-quality draft (Table S1). Comparative genome sequence analysis with *B. anthracis* Ames Ancestor revealed a single SNP in gene GBAA_2291, a sensor histidine kinase (position 2,139,247 in *B. anthracis* Ames Ancestor). This nonsynonymous SNP (TGG→TAG; W→stop) prevents the synthesis of a fully functional protein by generating an amber stop codon and resulting in the truncation of 149 aa in the predicted product (Fig. S2*A*). In the assembled sequence, the SNP position was covered by nine agreeing high-quality (median of 39.11) reads (Fig. S2*B*). By using PCR/sequencing-based assays designed to amplify a 533-bp amplicon, a 264-bp deletion was observed in a second Leahy letter isolate phenotypically similar to morphotype C. Because the genetic variation underlying the observed phenotype was different, this isolate was renamed morphotype D (LL7) (Table S3 and S4). The deletion was in-frame and should result in a protein that is 88 aa shorter (Fig. 3 *Upper Left*). The morphotype C point mutation location is encompassed within the morphotype D deletion (Fig. S2*A*). Detailed sequence analysis in *B. anthracis* Ames Ancestor demonstrated that two copies of a 10-bp repeat were present in

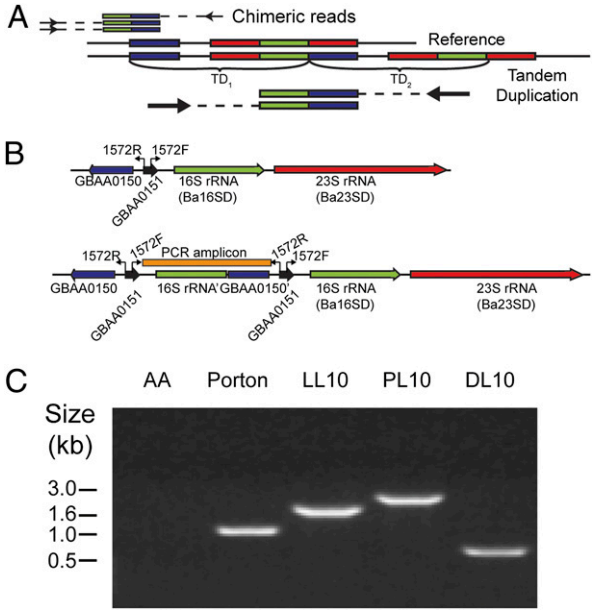

Fig. 2. Genomic characterization of morphotype A. (*A*) No mutation was originally discovered in morphotype A sequencing until a closer examination of the assembly revealed a number of chimeric reads and read pairs that disagreed with the structure of the original assembly, suggesting that a tandem duplication could have occurred. (*B*) PCR/sequencing assay that was designed to identify other mutations at this locus. The two primers, 1572F and 1572R, are both located within the coding region GBAA_0151, and a PCR amplicon is produced only if the coding region is at least duplicated (*Lower*). (*C*) Result of gel electrophoresis of the amplicons for this assay. The genomes amplified are *B. anthracis* Ames Ancestor (AA), *B. anthracis* Ames Porton (Porton; not previously known to have this duplication), *B. anthracis* Ames Leahy letter (LL10), *B. anthracis* Ames NY Post letter (PL10), and *B. anthracis* Ames Daschle letter (DL10; triplication of 822-bp amplicon).

GBAA_2291 and flanked the deleted region. A single copy of the repeat sequence was found in morphotype D, suggesting that recombination between these two repeats as a probable mechanism for the deletion (Fig. 3).

Morphotype E (Opaque). Morphotype E exhibited different phenotypic characteristics compared with the other variants in that colonies were smaller and more compact. Although the oligosporogenic phenotype was maintained, morphotype E displayed a temperature-sensitive, sporulation-deficient phenotype (Fig. 1 and Table 1). Comparative sequence analysis of morphotype E (LL18) with *B. anthracis* Ames Ancestor revealed no genetic variation of the chromosome, but a 21-bp in-frame deletion in a plasmid pXO1–encoded gene was identified. This locus, GBAA_pXO1_0205, was annotated as a tetratricopeptide repeat domain protein and putative response regulator. The sequence surrounding the deletion is of high quality with an average coverage of 9× (Fig. 3 *Upper Right*). As with the other morphotypes, a PCR/sequencing-based assay was designed (Table S3 and S4) that amplified a 503-bp amplicon surrounding the deletion. When applied to a second Leahy letter morphotype E isolate (LL19), the assay showed that LL19 carried a 9-bp in-frame deletion that was contained within the LL18 deletion (Fig. S3). One of the morphotype E genotypes containing the 9- or 21-bp deletion was identified in each of the evidentiary samples (Table S5). All mutants identified at this location by PCR were verified by sequence analysis.

Screening of *B. anthracis* Ames Isolates Relevant to the Amerithrax Investigation. Single-colony isolates recovered from several environmental locations relevant to the investigation—including the AMI building in Boca Raton, the NY Post office, a Trenton Post Office air filter, the Trenton National Institute for Occupational Safety and Health, Nguyen's house in New York City, Lundgren's house in Connecticut, Tom Brokaw's office, and NY Post Blvd—were screened with the PCR/resequencing assays described above. None of these isolates were positive for any of the morphotype assays, indicating that the identified mutations were not prevalent in the environment, even those associated with the investigation (Table S5); however, only environmental isolates were examined, and not populations, so the possibility that these mutations were present does exist. Positive controls included each of the different morphotypes isolated from each letter and negative control, including *B. anthracis* Kruger B, *B. anthracis* Western North America, *B. anthracis* CNEVA, *B. anthracis* Ames Ancestor, *B. anthracis* A1055, and *B. anthracis* Ames Florida, for which a genome sequence was available and predicted absence of these markers. As described above, *B. anthracis* Ames Porton was positive for morphotype A, with a repeat size of 1,260 bp.

Discussion

The forensic investigation that followed the 2001 mail attacks in Boca Raton, Washington, DC, and New York City required a multidisciplinary scientific team to characterize the evidentiary spores. The ultimate goal of this group was to determine the origin of the spores and potentially identify the perpetrator(s). The accuracy and reliability of whole-genome sequencing as a microbial forensic technique was not appreciated before the initiation of this project. This study provided the Amerithrax investigation with a set of validated genetic markers that were used to develop high-throughput quantitative PCR screening assays in an attempt to identify the potential source of the spores used in the mailings. These markers are specific for morphological variants identified

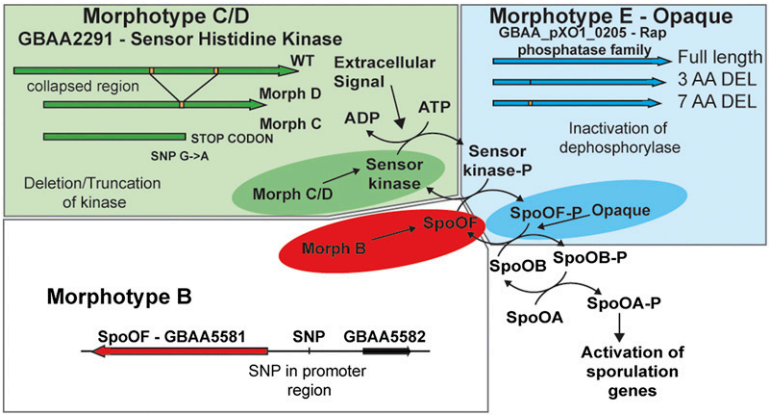

Fig. 3. Genomic characterization of morphotypes B, C/D, and E. Each of the loci impacted in these three morphotypes played a role in the sporulation cascade of *B. anthracis* Ames. Morphotype B (*Lower Left*) is characterized by an SNP in the promoter region of *spoOF*, one of the major phosphor-relay proteins in the sporulation cascade. Morphotypes C/D (*Upper Left*, green) were characterized by either an in-frame deletion (Morph D) or a truncation of the gene product resulting from an SNP in the sensor histidine kinase (GBAA_2291). Morphotype E (Opaque) (*Upper Right*, blue) is characterized by a 9- or 21-bp deletion in the pXO1 plasmid–encoded Rap phosphatase family protein (GBAA_pXO1_0205).

in spore preparations sent through the mail in the fall of 2001 (7, 8). The specific combination of morphological variants and associated sequence found in the letters' spore preparations linked these evidentiary samples to one another because they are not found in other samples (Table S5). These genotypes were later used to develop highly sensitive assays to screen a repository of *B. anthracis* Ames collected during the FBI investigation to attempt to identify potential biological source(s) of the spores sent in the letters. It is expected that the inclusion of multiple markers would provide greater resolving power than the use of a single genetic variant. Although this study highlights the power of whole-genome sequencing, the extensive microbiological analyses were key to the identification of the morphological variants.

Our study contributed to the development of genetic markers with exquisite specificity to support the Amerithrax investigation. For example, although current genotyping systems are able to differentiate strains of *B. anthracis*, they are unable to distinguish rare subpopulations within individual cultures. *B. anthracis* is highly monomorphic, and accurate genotyping relies on a combination of canonical SNPs and multiple-locus variable-number tandem repeat analyses (1, 4, 24–26). We started with the naïve approach that a single reference genome and sample would be sufficient for comparison. Read et al. (9) identified a number of polymorphic sites when comparing the Ames Porton and Ames Florida genomes; however, with only two points of reference, very little could be definitively ascertained. The comparison of the Ames Ancestor genome sequence identified no differences compared with that of the Ames Florida isolate, an isolate obtained directly from the individual infected at the AMI building (9). This finding suggests that the variation in the Ames Porton isolate was most likely caused by years of laboratory culture and treatments to remove the plasmids (18).

As part of the investigation, we sequenced the genomes of wild-type isolates from the spore preparations in each of the attack letters, with the exception of the Daschle letter (Table S1). These isolates could not be distinguished from *B. anthracis* Ames Ancestor based on morphological criteria or genome sequence, and they represented the numerical majority of the colonies in the spore preparations. Potentially, a maximum of 21 y of laboratory growth separated these isolates (Ames Ancestor was isolated in 1981), and no SNPs, INDELs, or rearrangements were observed. The chromosome and two plasmid sequences were identical.

The isolated morphotypes showed very few genetic variations compared with *B. anthracis* Ames Ancestor (Table S6). Each class of morphological variant was linked to a distinct genetic alteration. The genotype of three of the four morphological variants indicated transcriptional, translation, or posttranscriptional modifications of Spo0F, an integral protein in the early steps of the sporulation cascade (Fig. 3); these observations are in line with the observed phenotypes. Additionally, these findings are congruent with phenotypic observations of *spoOA* mutants, which are also defective in early sporulation steps (21).

Morphotype A. The extensive variation observed at this locus suggests that this is a hot spot for mutation. When examined at the molecular level, these isolates each contained a unique molecular structure at a single locus. Thus, the consistency of the phenotype could be attributed to the fact that each morphotype A isolate examined displayed genetic variation at the same locus, resulting in the same phenotype. A possible gene-dosage effect might also contribute to the phenotype because gene GBAA_0151, a hypothetical protein, was at least duplicated in each morphotype A isolate examined. Although the exact function of this gene is unknown, this study provides evidence that this locus plays a functional role in the sporulation cascade of *B. anthracis* Ames.

Morphotype B. This class of morphological variants contained a single SNP located in an intergenic region that appears to have a direct effect on the efficiency of sporulation in these morphological variants. This particular SNP, in the *spo0F* upstream region, may interfere with transcription, resulting in defect in the initiation of sporulation and yielding the observed oligosporogenic phenotype. The Spo0F protein has been identified in both *Bacillus subtilis* and *B. anthracis* as being involved in one of the early/intermediate phosphor-relay events that leads to sporulation (27) (Fig. 3), and mutants in *spo0F* have been generated that do not sporulate because of lack of signal (28, 29).

Morphotypes C/D. Morphotypes C/D are a family of variants identified initially in the Leahy letter spore preparations with mutations in the sensor histidine kinase GBAA_2291, which plays a role in the phosphorylation of Spo0F in the initial steps of the sporulation cascade (Fig. 3). Brunsing et al. (30) demonstrated that *B. anthracis* harbors a number of potential kinases, similar to KinA from *B. subtilis*, that play active roles in the phosphorylation of Spo0F and subsequently activate sporulation. Both of the genotypes (C and D) result in truncated forms of this sensor histidine kinase, suggesting a direct link between these mutations and the observed phenotypes (Fig. 1).

Morphotype E (Opaque). As we observed with morphotypes A and C/D, isolates with the morphotype E phenotype exhibited more than one genotype. We identified an in-frame deletion of 9 or 21 bp in the putative response regulator gene GBAA_pXO1_0205 on plasmid pXO1, resulting in the deletion of 3 or 7 aa. The observed phenotype suggested that these variations alter the peptide structure and/or function because it is the only genetic difference found on the genome of either morphotype E isolate. Bongiorni et al. (31) demonstrated that this particular response regulator is a member of the Rap phosphatase family, responsible for the dephosphorylation of Spo0F during sporulation (32, 33) (Fig. 3). Clearly the in-frame deletion within this protein is functionally deleterious. Further studies on the structure and function of this gene will be required to functionally characterize the phenotype.

The results presented here support the concept, among others, that laboratory growth of bacterial cultures may generate low-level genetic variation. Thus, the genomic DNA isolated for sequencing from a single culture could represent a mixture of the population members, abundant and not abundant. The underlying genetic characteristics of these variants could represent unique features of the originating microbial evidentiary samples. DNA assembly algorithms assume that the majority of the input sequence is from a single clonal source and formats the resulting data into a consensus sequence where the majority of the sequence agrees and represent the dominant, most abundant member present depending on genome sequence coverage (34, 35). The subpopulations containing the morphotypes described in this study were present in the letter spore samples. However, initial analysis of the population culture by Sanger sequencing failed to identify them because they represented an extremely small fraction of the population used to generate the genomic DNA for sequencing and the genome sequence coverage was between 8× and 12×. Some would argue that next-generation sequencing technologies that afford increased genome sequence coverage would now be a better choice for these applications. However, with the increased error rate and lack of accurate associated quality scores, these methods may not have been able to identify the signal from the noise and identify from a single culture both the wild type and the morphological variants. Nevertheless, population genomics of bacterial cultures will need to be further investigated with these newer sequencing technologies and better assembly algorithms.

Given the relative time and cost of whole-genome sequencing, selecting the isolates to sequence is a critical issue. The forensic nature of any future investigation is likely to be unique, and making generalizations about rules for selection is difficult; however, genome sequences from diverse type strains of all potential biothreat microbial species will be essential for future microbial forensic investigations, mainly as high-quality refer-

ences. With these key references in hand, we suggest that evidentiary samples could be sequenced to draft quality and examined for variations by comparative analysis to the high-quality references. In this study, three genomes were closed; however, no novel high-quality sequence variation was discovered in the closed genomes that was not identified in the draft sequence data. A draft genome can be rapidly obtained at a much lower cost and should be analyzed before the decision is made to complete the genome. The caveat to using draft sequences is that there *must* be a high-quality closed reference genome already available. In this particular study, it may have been more economical to sequence to draft quality one of each of the morphotypes, analyze the data for genetic variation, and then test the other isolates of the same morphotype for the presence of these sequence variants. This was the approach that was used later in the project for morphotypes C/D and E from the Daschle letter, and it proved to be an efficient and cost-effective approach to validate these variable sites.

The use of *B. anthracis* in this bioterror attack taught us important lessons about the integration of whole-genome sequencing for forensic applications; however, it is unclear if the same approaches could be taken with other biological agents. *B. anthracis* is a young and genetically homogenous species that has features that contribute to this reduced genetic variability, such as spore formation and dormancy. Spores can remain dormant for extended periods of time (decades or perhaps centuries) and, during this time, should not accumulate genetic mutations, thus providing a snapshot of the source material. If a similar attack had been perpetrated with *Escherichia coli*, *Salmonella*, or any non-spore-forming bacteria, it would have been more difficult to attribute minor genetic variations to a source sample. However, a perfect genomic match between source and evidence would be considered a strong match and would be unlikely to be made with other isolates. The signal-to-noise ratio may be too great in other species and accurate attribution too onerous of a task with genome sequencing.

This study highlights the resolving power of whole-genome sequencing and careful comparative genomic analyses. We demonstrate that it is possible to identify a single nucleotide change in a genome that contains more than 5 million bp and attribute a visible phenotype to that single substitution. When applied to microbial forensics, bacterial whole-genome sequence data appear to be the ultimate evidence; much like a human genetic fingerprint, it could link microbial evidence to its source(s). This study is an example of the development of forensic markers for a biological weapon by whole-genome sequence comparison. These markers are stable, varied in genomic location, and unique compared with the Ames Ancestor genome sequence. These markers together constitute a molecular fingerprint that may not be found in any other spore preparation but the one from which the spores sent through the mail in the fall of 2001 originated (Fig. S4). These identifying markers were transferred to another scientific investigative team that developed high-throughput and sensitive molecular assays to screen for the presence of these morphological variants in DNA extracted from cultures of *B. anthracis* Ames collected during the FBI investigation. Using these assays, the investigative team was able to demonstrate that cultures within the FBI repository contained the four morphological variants together, which supported the FBI investigative efforts to identify the source of the 2001 letter spores.

Materials and Methods

Detailed materials and methods describing microbiological examination, comparative genome analysis, and morphotype PCR/resequencing assay development can be found in *SI Materials and Methods*.

ACKNOWLEDGMENTS. This work was jointly supported by federal funds from the Department of Justice under Contract J-FBI-02-016; the National Institute of Allergy and Infectious Diseases, National Institutes of Health, under Contract N01-AI15447; and National Science Foundation Small Grant for Exploratory Research MCB-202304.

1. Van Ert MN, et al. (2007) Global genetic population structure of *Bacillus anthracis*. *PLoS ONE* 2:e461.
2. Bergman NH, et al. (2006) Transcriptional profiling of the *Bacillus anthracis* life cycle in vitro and an implied model for regulation of spore formation. *J Bacteriol* 188: 6092–6100.
3. Keim P, et al. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *J Bacteriol* 182:2928–2936.
4. Keim P, et al. (1997) Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. *J Bacteriol* 179:818–824.
5. Harrell LJ, Andersen GL, Wilson KH (1995) Genetic variability of *Bacillus anthracis* and related species. *J Clin Microbiol* 33:1847–1850.
6. Pearson T, et al. (2004) Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA* 101:13536–13541.
7. Jernigan DB, et al. (2002) National Anthrax Epidemiologic Investigation Team (2002) Investigation of bioterrorism-related anthrax, United States, 2001: Epidemiologic findings. *Emerg Infect Dis* 8:1019–1028.
8. Jernigan JA, et al.; Anthrax Bioterrorism Investigation Team (2001) Bioterrorism-related inhalational anthrax: The first 10 cases reported in the United States. *Emerg Infect Dis* 7:933–944.
9. Read TD, et al. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296:2028–2033.
10. Jackson PJ, et al. (1997) Characterization of the variable-number tandem repeats in *vrrA* from different *Bacillus anthracis* isolates. *Appl Environ Microbiol* 63:1400–1405.
11. Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T (2002) Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. *Emerg Infect Dis* 8:1111–1116.
12. Lindstedt BA (2005) Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* 26:2567–2582.
13. Tettelin H (2009) The bacterial pan-genome and reverse vaccinology. *Genome Dyn* 6: 35–47.
14. Harris SR, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474.
15. Read TD, et al. (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* 423:81–86.
16. Green BD, Battisti L, Koehler TM, Thorne CB, Ivins BE (1985) Demonstration of a capsule plasmid in *Bacillus anthracis*. *Infect Immun* 49:291–297.
17. Okinaka R, et al. (1999) Sequence, assembly and analysis of pX01 and pX02. *J Appl Microbiol* 87:261–262.
18. Okinaka RT, et al. (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J Bacteriol* 181:6509–6515.
19. Ravel J, et al. (2009) The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". *J Bacteriol* 191:445–446.
20. Budowle B, et al. (2005) Genetic analysis and attribution of microbial forensics evidence. *Crit Rev Microbiol* 31:233–254.
21. Worsham PL, Sowers MR (1999) Isolation of an asporogenic (spoOA) protective antigen-producing strain of *Bacillus anthracis*. *Can J Microbiol* 45:1–8.
22. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175–185.
23. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194.
24. Keim P, et al. (2004) Anthrax molecular epidemiology and forensics: Using the appropriate marker for different evolutionary scales. *Infect Genet Evol* 4:205–213.
25. Van Ert MN, et al. (2007) Strain-specific single-nucleotide polymorphism assays for the *Bacillus anthracis* Ames strain. *J Clin Microbiol* 45:47–53.
26. Kenefic LJ, et al. (2008) Texas isolates closely related to *Bacillus anthracis* Ames. *Emerg Infect Dis* 14:1494–1496.
27. Burbulys D, Trach KA, Hoch JA (1991) Initiation of sporulation in *B. subtilis* is controlled by a multicomponent phosphorelay. *Cell* 64:545–552.
28. Eichenberger P, et al. (2004) The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biol* 2:e328.
29. Sonenshein AL (2000) Control of sporulation initiation in *Bacillus subtilis*. *Curr Opin Microbiol* 3:561–566.
30. Brunsing RL, et al. (2005) Characterization of sporulation histidine kinases of *Bacillus anthracis*. *J Bacteriol* 187:6972–6981.
31. Bongiorni C, Stoessel R, Shoemaker D, Perego M (2006) Rap phosphatase of virulence plasmid pXO1 inhibits *Bacillus anthracis* sporulation. *J Bacteriol* 188:487–498.
32. Perego M (2001) A new family of aspartyl phosphate phosphatases targeting the sporulation transcription factor Spo0A of *Bacillus subtilis*. *Mol Microbiol* 42:133–143.
33. Perego M, Brannigan JA (2001) Pentapeptide regulation of aspartyl-phosphate phosphatases. *Peptides* 22:1541–1547.
34. Myers EW, et al. (2000) A whole-genome assembly of *Drosophila*. *Science* 287: 2196–2204.
35. Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. *Brief Bioinform* 5:237–248.