



**DEPARTMENT OF DEFENSE  
OFFICE OF FREEDOM OF INFORMATION  
1155 DEFENSE PENTAGON  
WASHINGTON, DC 20301-1155**

JAN 28 2011

Ref: 10-F-1062

Mr. Steven Aftergood  
Federation of American Scientists  
1725 DeSales Street NW, Suite 600  
Washington, DC 20036

Dear Mr. Aftergood:

This responds to your November 14, 2009, Freedom of Information Act (FOIA) request filed with the Office of the Director of National Intelligence. You requested a copy of the following report prepared for the National Counterproliferation Center by the JASON advisory panel: "Microbial Forensics" by T. Stearns, et al, JASON Report No. JSR-08-512, 2008. On May 25, 2010, ODNI referred your request along with the responsive document to this office for processing and direct response to you. The Office of the Secretary of Defense/Joint Staff determined the enclosed document can be released without excision. This action closes your request in this office.

Sincerely,

*Nicholas A. Cormier*

*for* Paul J. Jacobsmeyer  
Chief

Enclosure(s):  
As stated

~~FOR OFFICIAL USE ONLY~~

---

## Microbial Forensics

---

May 2009

JSR-08-512

~~Distribution authorized to US Government & Contractors; Administrative/Operational Use; May 9, 2009.  
Other requests for this document shall be referred to National Counterproliferation Center.~~

JASON  
The MITRE Corporation  
7515 Colshire Drive  
McLean, Virginia 22102-7508  
(703) 983-6997

~~FOR OFFICIAL USE ONLY~~

## Contents

<b>1 EXECUTIVE SUMMARY</b>	<b>1</b>
<b>2 INTRODUCTION</b>	<b>5</b>
2.1 Scope of the Study.....	6
2.2 Tularemia and <i>F. tularensis</i> .....	7
2.3 <i>F. tularensis</i> Genetics and Genomics.....	10
2.4 Federal Microbial Forensics Infrastructure .....	11
<b>3 METHODS AND CRITERIA FOR DETERMINING GENETIC RELATEDNESS</b>	<b>13</b>
3.1 The Implications of Mutation Rate and Clonal Growth for Microbial Forensics .....	13
3.2 DNA Polymorphisms, DNA Sequencing and Microbial Forensics .....	15
3.3 Laboratory-based Experiments on Genetic Change vs. Natural Variation.....	18
<b>4 APPLICATION TO THE ATTACK SCENARIO</b>	<b>21</b>
<b>5 CONCLUSIONS AND RECOMMENDATIONS</b>	<b>25</b>
<b>6 REFERENCES</b>	<b>27</b>

## 1 EXECUTIVE SUMMARY

### *Introduction*

Microbial forensics is an emerging field devoted to the development of methods to characterize microbial samples for the purpose of comparative analysis in support of investigations into criminal or terrorist acts. The comparison and classification of microbial isolates has been a focus of microbiological studies for more than 100 years. Historically, such comparisons have relied on overt characteristics of microbes, such as cell morphology, Gram staining, serology, and growth on diagnostic media. The advent of DNA sequence-based methods has dramatically improved our ability to discern relatedness between two isolates, and offers the resolution required for attribution in forensic investigations. These same sequence-based methods of comparison are widely used in academic research in the fields of microbial evolution, taxonomy, and epidemiology. However, the needs of microbial forensics differ, in that the results may be used in criminal prosecution or in making national security decisions, and therefore greater resolution and certainty are required.

### *Study Charge*

JASON was asked to address the development of a research roadmap that would provide an underpinning for improved microbial forensic capabilities. The task scenario focused on *Francisella tularensis*, the bacterial agent that causes tularemia. *F. tularensis* is a Select Agent, and considered to be a serious bioterrorism threat. In the scenario, a biological attack is carried out in three U.S. cities over the course of one week, with samples available from victims at all three attack sites, from patients, from a dissemination device found at one of the sites, and from two sources collected by the Intelligence Community. The task was to determine the current state of genetic knowledge regarding *F. tularensis*, to identify gaps in that knowledge relevant to microbial forensics capabilities, and to propose a research plan that describes a scientific path forward. Although *F. tularensis* was the task scenario organism, the research plan should be one that allows for the broader microbial forensics capabilities that are needed.

### *Summary*

JASON considered the following specific areas:

- 1) Methods and criteria for determining genetic relatedness

**FOR OFFICIAL USE ONLY**

- 2) Laboratory-based experiments on genetic change vs. natural variation
- 3) Metagenomics opportunities and concerns
- 4) Potential for geolocation

The technologies for assessing genetic relatedness all depend on differences in DNA sequence between individual organisms. It is important to recognize that microbial DNA forensics is fundamentally different from human DNA forensics in that most microbes grow by asexual reproduction, and most are capable of a many-fold increase in cell number in a short time. This potentially results in a large number of individuals that are genetically identical. In contrast, every human being is genetically distinct (with the exception of identical twins), and can be distinguished by assaying a small number of sites in the genome.

The technologies in current use for genetic analysis in microbial forensics are of two sorts: 1) assays in which known, genetically polymorphic sites are probed to determine the alleles present; and 2) whole genome sequencing, which does not depend upon prior knowledge. The existing polymorphism assays differ in their level of resolution and in the rate of change of the specific polymorphism assayed, but they have in common that they are rapid, inexpensive, and scalable to be high throughput. These assays are useful in situations in which rapid identification of species and strain is required, for example, during the early stages of a natural outbreak or terrorist attack. However, whole genome sequence is vastly superior in information content, and should be the method of choice when samples of forensic interest have been identified. The cost in time and money of sequencing microbial genomes was a concern in the past, but next-generation sequencing technologies are emerging that generates very large amounts of sequence data in a single run. These technologies are fundamentally different from traditional Sanger sequencing in that the output is very many short-sequence reads that can provide greater than 30-fold sequence coverage of the genome per run for a typical bacterial genome. The immediate consequences of this new sequencing power are reflected in many of our conclusions and recommendations.

The primary conclusions of the study are that:

- 1) Microbial forensics is a powerful tool, but is most useful as a complement to field work, human intelligence, evidence gathering, and other traditional forms of forensics.
- 2) Whole-genome DNA sequencing provides the ultimate level of genetic resolution for forensics and will replace the current forensic methods of analyzing DNA

~~FOR OFFICIAL USE ONLY~~

polymorphisms. It is also revolutionizing metagenomics. The microbial forensics community should be at the forefront of this transition.

- 3) A quantitative assessment of the meaningfulness of a “match” for forensics is difficult, because of the asexual growth of bacteria, lack of information about population structure, mutation rate, and number of generations that separate two samples, and the unknown effects of human intervention.
- 4) Quality of relatedness information (phylogenetic trees) is limited by the quantity, quality, and diversity of the data collected.
- 5) Although DNA-based methods of strain comparison are the most powerful, methods that exploit other signatures, such as gene expression and isotopic composition will likely be important forensically.
- 6) Improving microbial forensics capabilities also improves capabilities for analyzing natural outbreaks, and *vice versa*.

The chief recommendations of the study are the following:

- 1) For all relevant pathogens, the National Biodefense Forensic Analysis Center (NBFAC) should sample the diversity of both natural and laboratory strains, and sequence at low fidelity to get a sense of sequence variation. A few representative strains should be chosen for high-fidelity sequencing to establish “gold standard” sequences for each pathogen. High-fidelity sequence of a related outgroup should also be acquired (*Bacillus cereus* for *Bacillus anthracis*, for example).
- 2) The broad group of government agencies with an interest, including NBFAC, IC, FBI, CDC, and USDA, should develop and promote best practices for microbial sample collection that best preserve the genetic and non-genetic signatures of interest.
- 3) NBFAC, with its bioinformatics partners at LLNL, should exploit data from emerging metagenomic projects to expand knowledge of natural diversity. Bioinformatics tools better able to deal with this growing sequence database are needed. Some of these are currently being developed by the academic community, others by the commercial vendors of sequencing technology, but some may need to be developed in-house.
- 4) The strain collection efforts of NBFAC and the National Bioforensic Reference Collection should be expanded, with greater support, and coupled with the development of a relational database that includes genetic, geographic, and other contextual information. Sequences themselves, in addition to organisms, should be considered part of this database.

**~~FOR OFFICIAL USE ONLY~~**

JASON also identified two longer-term basic research goals that will impinge upon microbial forensics capabilities:

- 1) Develop capabilities for novel microbial forensics methods, such as transcriptome analysis, characterization of epigenetic modifications, and metabolomics.
- 2) Develop the “pathogenome” (the molecular and genetic mechanisms of pathogenicity) as an organizing concept for microbial pathogens, complementing “species” organization.

This study provided an opportunity to consider microbial genetics and population biology issues in the context of forensics, taking into account emerging technology that is changing the field. Studies of this kind will be useful in the future to keep abreast of how ongoing developments in science and technology affect changing microbial forensics needs.

## 2 INTRODUCTION

In this report we summarize the considerations and conclusions of the 2008 JASON Summer study on Microbial Forensics. The charge to JASON, from the National Counterproliferation Center (NCPC) was to:

Develop a research roadmap to address and support criteria for 'match' determination as it pertains to genetic sequence variation comparisons for microbiological samples that would be compared as part of an attribution investigation.

- Develop scientifically defensible standards
- Account for differences between species and within species
- Develop assurance and validation criteria
- Identify the types of bioinformatics tools needed
- Describe ways in which to best leverage current and future technological advances to support forensics.

The specific context in which the above issues were considered was a forensic scenario provided by the sponsor:

A covert biological attack is carried out in 3 separate US cities with no claims of responsibility over a period of 1 week. The attack in each case involved the covert release of a biological aerosol in a venue where a large number of people were either present or transiting through. The first evidence of an attack is derived by the appearance of patients with severe pulmonary infections and the etiologic agent is identified in clinical laboratories as *Francisella tularensis* type A. Approximately 5% of the cases result in fatalities. Epidemiological investigation identifies the likely sites as a shopping mall, a train station, and a movie theater. The FBI and other federal, state, and local agencies respond quickly to these suspected sites, and a large number of samples are taken. In one instance a crude improvised dissemination device is recovered from an air duct at the movie theater; the device had some residual material in it, but went undiscovered for several days following the attack. In addition to the environmental samples taken in each venue, the clinical samples recovered from the victims, and the residual material present in the movie theater device, the FBI learns that the Intelligence Community had several instances in which *F. tularensis* type A DNA signatures had been collected from overseas exploitation operations. One of those operations involved a site exploitation of a bombed out terrorist laboratory following counter-terrorism operations

## FOR OFFICIAL USE ONLY

conducted by the U.S. military, and the second involved a freezer vial containing non-viable *F. tularensis* type A recovered from a suspect state-sponsored biological weapons program. (Figure 1 illustrates the types of samples available for comparison).

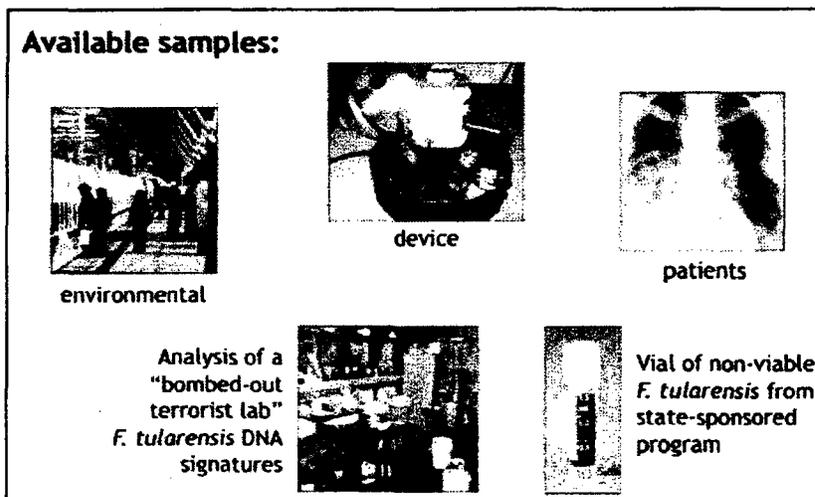


Figure 1. Samples of *Francisella tularensis* organisms or DNA available for analysis in the attack scenario.

### 2.1 Scope of the Study

Although many types of forensic evidence would be useful in a case such as that provided in the scenario, JASON was asked to focus specifically on issues pertaining to genetic analysis of samples. Questions that arise directly from the attack scenario, and are potentially addressable by genetic comparisons, include the following:

- 1) Was a single preparation of material used for all of the attack locations, or were multiple preparations made?
- 2) Can the samples from patients be related between attack sites, and between patients and dissemination device?
- 3) Can *F. tularensis* in environmental samples at the attack sites be distinguished from naturally-occurring *F. tularensis*?
- 4) Can samples of *F. tularensis* from domestic laboratories, or IC sources be definitively linked to the attack material, or excluded, based on genetic similarity?

~~FOR OFFICIAL USE ONLY~~

- 5) Is geolocation based on genetic sequence analysis feasible and could it be useful for attribution?

In the course of addressing these questions we considered the state of knowledge for the pathogen in question, *F. tularensis* type A, but also considered the requirements for microbial forensics as they relate to bacterial pathogens more broadly because many of the issues are species-independent. Although viral and fungal pathogens are also potentially important threat organisms, we were asked to focus on bacterial pathogens. We thank the following briefers for helpful insights and discussion:

**NCPC**

Paul Keim – Northern Arizona University, Flagstaff; TGen, Phoenix

John Taylor – University of California, Berkeley

Kostas Konstantinidis – Georgia Institute of Technology

James Burans – National Bioforensic Analysis Center

David Relman – Stanford University

**2.2 Tularemia and *F. tularensis***

Tularemia, popularly known as ‘rabbit fever,’ is an acute, febrile illness that results from infection by the contagious bacterium *Francisella tularensis*. Initial symptoms of disease typically appear in 3–5 days, but depending upon the route and severity of the infection, this incubation period can range from as little as 1 day to as many as 14 days. *F. tularensis* is the most infectious bacterial organism known: it has been estimated that just 10–50 bacterial cells are able to induce the life-threatening pulmonary form of tularemia. Under current law *F. tularensis* is classified as a CDC Select Agent.

The causative agent of tularemia was first isolated in 1911, in Tulare County CA, by noted bacteriologist Dr. Edward Francis, who traced an outbreak of fever to the local ground squirrel population. Tularemia is a zoonotic bacterial disease that infects not only humans, but also lagomorphs (rabbits, hares, and pikas), most common rodents (including squirrels, rats, voles, and mice), and some birds. Over 100 species of wild and domestic animals are known to harbor tularemia bacteria, including sheep, cats, skunks, raccoons, and even pet hamsters. In the U.S., major outbreaks of the disease have claimed commercially important numbers of sheep, mink, beaver, and foxes. The disease can also be spread by biting insect vectors, including ticks, fleas, deerflies, and mosquitoes. The animal reservoir (host) for *F. tularensis* is not presently

~~FOR OFFICIAL USE ONLY~~

established, but dogs and wolves appear to be fairly resistant to infection and may also serve as maintenance hosts for ticks.

Four major subspecies of *F. tularensis* are currently recognized. The subspecies are similar morphologically, biochemically, and genetically, but differ in their geographical distribution. *F. tularensis tularensis* is prevalent throughout North America, including 49 of the 50 U.S. states (Hawaii being the sole exception), and is classified as the Type A strain, which is responsible for the most virulent form of the disease. The three Type B subspecies produce a milder form of disease that is rarely life-threatening in humans. These include: *F.t. holartica*, found throughout the Northern Hemisphere and the predominant form in both Europe and Asia; *F.t. mediasiatica*, found in some Central Asian Republics; and *F.t. novicida*, which is a comparatively rare strain found mainly in North America, and more recently in Australia. The weakly virulent *F.t. novicida* strain has been used as a basis for a tularemia vaccine, but that vaccine is in very limited use and restricted to health care workers and others at specific risk for disease. No vaccine is currently available to the public.

Widespread distribution of a vaccine against tularemia is not thought to be necessary. Despite the highly infectious nature of Type A strains, the disease is not common in the U.S. human population. Throughout the 1930's and 1940's, 2,000–3,000 cases per year were reported, and many of these were attributable to eating contaminated meat, coming mainly from the hunting of diseased squirrels and rabbits during the Depression era. With the subsequent reduction in such hunting, the incidence of tularemia decreased markedly, with only 1,368 cases reported during the decade of 1990–2000. Sixty percent of those cases came from just four states where squirrels and rabbits are still hunted (AR, MO, SD, and OK). Today, tularemia is rarely fatal when promptly diagnosed because the causative agent is susceptible to a wide variety of antibiotics. Given its low incidence and mortality rates, tularemia was removed from CDC "reportable disease" status in 1994. However, it was reinstated as a reportable disease in 2000 because of continuing concerns that *F. tularensis* might be used as an aerosolized bioweapon.

Arguably, *F. tularensis* may have been the first bioweapon to be used in recorded history. The ancient Hittites, who lived in portions of what is today Syria and Turkey, were reported to have supplied "cursed rams" carrying disease to their enemies, the Arzawans, in the year 1,320 B.C.E. By some accounts, the circumstances and symptoms were consistent with the ingestional form of tularemia. More recently, tularemia bacteria were weaponized by BW programs in both the former Soviet Union and the United States. Both sides developed imperfect tularemia vaccines as well, which tended to produce adverse reactions and were not well-suited for general use.

## ~~FOR OFFICIAL USE ONLY~~

According to Ken Alibek (Kanatjan Alibekov), a defector who served as the First Deputy Director of Biopreparat, the Soviet program had worked to genetically engineer a "triple-resistant" strain of *Francisella* during the 1970's and 1980's that would retain wild-type levels of infectivity while evading a spectrum of antibiotics, but was unsuccessful. It is reported that the U.S. stocks of weaponized *F. tularensis* were destroyed in 1973 and the former Soviet program was abandoned in 1992.

Tularemia is notable for the many ways that one can get the disease, as well as the wide variety of forms that it can take. *Francisella* infection can occur across the skin or mucous membranes, through minor cuts or abrasions, or via the bite of an infected insect or rodent. Infection by this route causes the *ulceroglandular* form of the disease, which accounts for 75–85% of all reported cases. Untreated, ulceroglandular tularemia has a 5–7% mortality rate. The overall case mortality rate for all forms in the U.S., with treatment, was just 1.4% between 1985 and 1992. The second-most common form of tularemia is the pulmonary, or *pneumonaic* form. Approximately 30% of natural cases are estimated to be of this type, and it is also the form that would be produced by a deliberately aerosolized BW agent. Pulmonary tularemia is the most lethal form of the disease, with a mortality rate of close to 50%. In nature, the pulmonary form can be acquired by the inhalation of aerosolized particles bearing live bacteria, such as those produced by the handling of contaminated rabbit pelts, spoor from infected animals, or even soil or road dust. An outbreak of tularemia in 2000 in Martha's Vineyard that produced 15 cases of pneumonaic tularemia with one fatality (and has had recurrences every summer since that time) has been attributed, without direct evidence, to the exposure of landscapers to aerosolized lawn clippings.

Tularemia may be acquired by ingesting contaminated meat or water, where it causes the so-called *typhoidal* form. Tularemia bacteria represent one of only two Select Agents known to maintain virulence in water, the other being anthrax spores. Typhoidal tularemia is fairly uncommon, but the mortality rate is high, similar to that for pulmonary tularemia.

Pre-9/11, the CDC estimated (in 1997) that the cost of a deliberate attack on the U.S. with aerosolized tularemia might be \$5.4B per 100,000 people infected. However, improved methods for diagnosis, general awareness of BW threats, and the numerous countermeasures taken by the U.S. government since the anthrax letters attacks during Fall 2001 have all likely reduced that number substantially. Still, given its natural properties, *F. tularensis* remains a leading candidate for an agent to be used in a BW attack. Finally, because *F. tularensis* type A is an endemic species in the U.S., with a broad distribution, it poses significant problems for the current

generation of biodetectors that attempt to score the presence of aerosolized bacteria. The publicized cases of these ultra-sensitive detectors detecting *F. tularensis* attest to the ubiquity of this organism in our natural environment.

### 2.3 *F. tularensis* Genetics and Genomics

The complete genome sequence of *F. tularensis* strain Schu S4 has been determined (Larsson et al., 2005). The genome is a 1.89 Mb circular chromosome with approximately 1,800 predicted protein coding sequences. Like most bacterial genomes, the *F. tularensis* genome contains many transposable elements, including 50 copies of a transposon belonging to the Tc-1 mariner family. Interestingly, this transposon is usually found in eukaryotes, and it is possible that it was acquired from an insect host.

The genome sequence, as well as previous 16S RNA sequence analysis, suggests that *F. tularensis* has no close relatives among known pathogens. Figure 2 shows a phylogenetic tree including *F. tularensis* and other bacterial pathogens.

The virulence mechanism of *F. tularensis* is not well-understood, but it is known that genes within a pathogenicity island relevant to growth within macrophage cells are regulated by the transcription factor MglA (Lauriano et al., 2004). Indeed, microarray analysis of genes regulated by MglA has recently identified five previously uncharacterized virulence factors (Brotcke et al., 2006).

A second whole genome sequence of an *F. tularensis* isolate was recently reported (Beckstrom-Sternberg et al., 2007). The *F. tularensis* type A subspecies is composed of two clades, A.I and A.II. There is some geographical separation between the two clades, with A.I being found

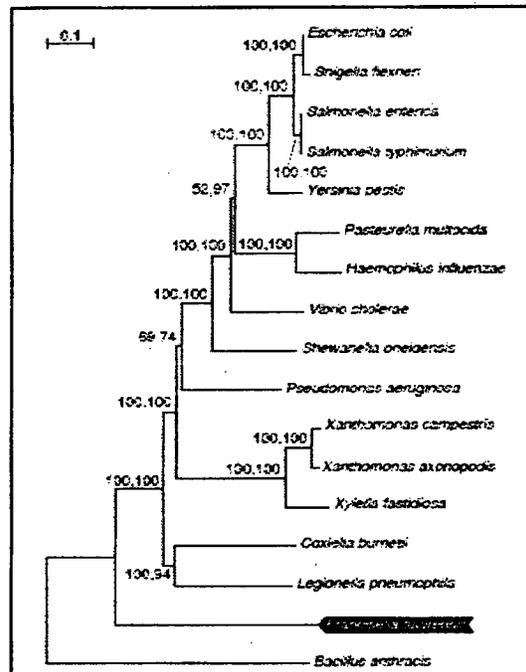


Figure 2. Phylogenetic relationship of 16  $\gamma$ -proteobacterial species inferred from a concatenated alignment of ten proteins. The topology, branch lengths and bootstrap support are according to the reconstruction with the neighbor-joining method. From Larsson et al. (2005).

predominantly in the southern U.S., and A.II in the mountainous western U.S. The Schu S4 strain sequenced by Larsson et al, (2005) is an A.I strain, whereas the newer sequence is from an A.II strain. There are more than 4,000 polymorphisms between the two genomes, as well as more substantial rearrangements, often flanked by transposable elements. Interestingly, a large number of transposon-flanked genome rearrangements were also noted in comparison of the Schu S4 sequence with that of a *F. tularensis* type B isolate (Petrosino et al., 2006). Although it is difficult to correlate estimates of divergence rate with real time, based on the degree of divergence it is clear that the two type A clades of *F. tularensis* separated from each other long ago.

#### **2.4 Federal Microbial Forensics Infrastructure**

A presidential directive dated April 28, 2004 named the National Bioforensic Analysis Center (NBFAC) as “the lead Federal facility to conduct and facilitate the technical forensic analysis and interpretation of materials from biocrime and bioterror”. NBFAC is one component of the National Biodefense Analysis and Countermeasures Center (NBACC), under the Department of Homeland Security. The other component is the Biological Threat Characterization Center (BTCC), which conducts studies and laboratory experiments to better understand biological threats. NBFAC has a “hub and spoke” organizational structure (Figure 3), with FBI and other customers providing samples for analysis, the NBFAC hub conducting casework, evaluating and validating new assays, and maintaining continuous operation. A variety of spoke affiliates provide surge capacity for casework, and increased analytic capabilities; these include Lawrence Livermore National Laboratory (LLNL) for bioinformatics, the Chemical and Biological Information Analysis Center (CBI) for mass-spectrometry characterization of chemical samples, and the Naval Medical Research Center (NMRC) for expertise on Rickettsial diseases. NBACC and NBFAC will be housed in a new building to be completed in early 2009. JASON had extensive discussions with James Burans, Ph.D., Director of NBFAC, and was convinced of the benefits of this integrated program for addressing bioforensic challenges.

~~FOR OFFICIAL USE ONLY~~

~~FOR OFFICIAL USE ONLY~~

### 3 METHODS AND CRITERIA FOR DETERMINING GENETIC RELATEDNESS

When a sample of interest is referred to NBFAC, a range of tools are applied to determine the nature of the organism in question (Figure 4). These include traditional microbiological assays, such as tests of ability to metabolize different sugars, immunological methods that recognize surface molecules, and electron microscopy to determine morphology (Figure 4). Each of these

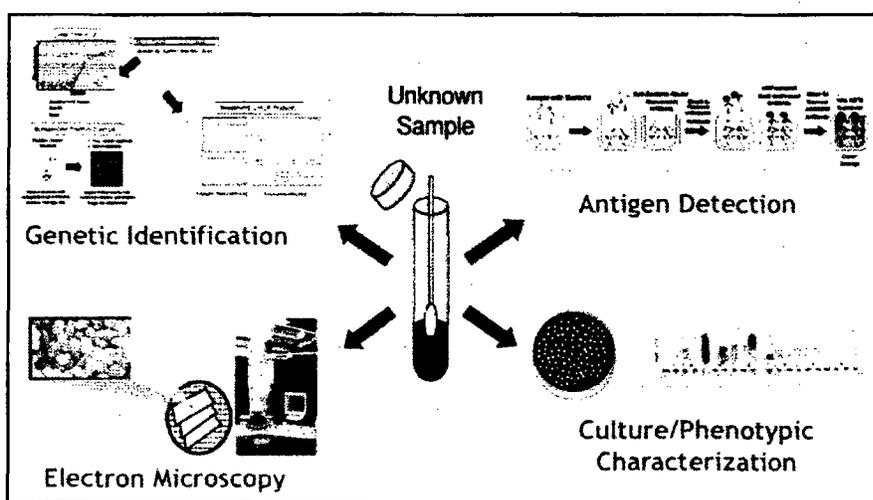


Figure 4. Tools used to characterize unknown microbial samples at NBFAC. Courtesy of James Burans.

tools has advantages, particularly when material is limited or when a rapid determination of species is needed. However, only genetic identification has the resolution required for forensic investigations in which the relationship between two isolates is the key issue, and we therefore focus on that method.

#### 3.1 The Implications of Mutation Rate and Clonal Growth for Microbial Forensics

All genetic methods of identification depend on nucleic acid sequence polymorphism between organisms. It is instructive to compare the nature and meaningfulness of genetic comparison for humans and bacteria. Knowledge of polymorphic loci in humans has led to powerful assays to identify individuals uniquely for forensic investigations. Given the number of humans on earth (approaching  $7 \times 10^9$ ), a relatively small number of polymorphic loci suffice for unique

~~FOR OFFICIAL USE ONLY~~

identification. The DNA Identification Act (1994) led to the establishment by the FBI of a Combined DNA Index System (CODIS) that is now used in the U.S. and more than 25 other countries (Baechtel et al., 1991). CODIS makes use of 13 loci in the human genome that can be distinguished by simple polymerase chain reaction (PCR) assays; the profile of alleles at these loci in different populations within the U.S. is known. The genetic diversity created by sexual reproduction in humans is such that, other forensics issues aside, assaying the alleles presents at the CODIS loci allows a unique match to be made between a forensic sample and an individual human. An obvious exception is identical twins, who would be indistinguishable by the CODIS assays.

We wish to stress that there is a fundamental difference between the use of genetic markers for forensics in humans compared to their use in microbial forensics. Most microbes reproduce asexually, so that vast numbers of "individuals" can be genetically identical, or nearly

identical. Figure 5 shows some of the relevant numbers to consider. The average bacterial genome consists of several megabases of DNA, and the mutation rate is approximately  $10^{-9}$ . Therefore, there are approximately 0.002 mutations per genome in a population, and most cells in a population are genetically identical. However, it is also important to realize that very large numbers of individual cells can be grown on solid medium in plates (approximately  $10^9$  cells per bacterial colony), or in liquid medium (approximately  $10^{12}$  cells per liter of culture). Given the population size that can be achieved, at least in the laboratory under optimum conditions, it is clear that there would be many variants — cells that differ genetically from the canonical individual — in a culture. Therefore, in an asexually growing population of bacterial cells, most individuals are genetically identical to each other, but there also are likely to be many variants, depending on the mutation rate and population size. Mutation rates for different species and for different types of mutations within a single species vary widely (Denamur and Matic, 2006), so the above is meant to be illustrative, rather than definitive.

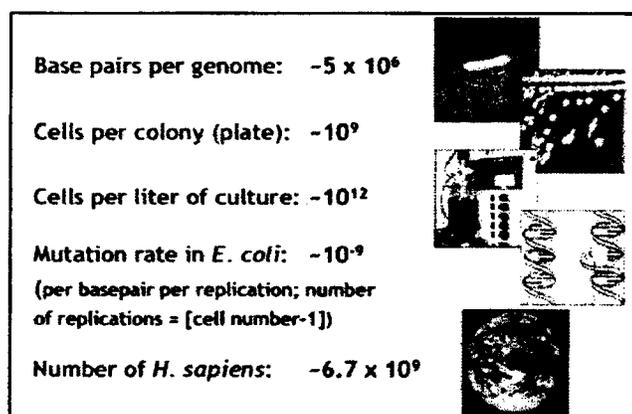


Figure 5. Reference numbers for microbial genetics.

### 3.2 DNA Polymorphisms, DNA Sequencing, and Microbial Forensics

DNA is used as a biometric because it is a relatively stable molecule and is information-rich, and because reliable and inexpensive methods have been developed that allow detailed analysis of very small amounts of material. In addition, there is a large and growing dataset of genome sequences from the “tree of life”, and of variants of those sequences found in natural and laboratory populations.

Typing organisms based on DNA requires DNA polymorphisms, or changes in DNA sequence between different individuals. There are four main types of DNA polymorphisms:

- 1) Single nucleotide polymorphisms (SNPs)
- 2) Insertions and deletions (indels)
- 3) Variations in length of repeated sequences (VNTRs, STRs)
- 4) Large-scale genome rearrangements (inversions, large insertions and deletions)

Although these DNA polymorphisms have been recognized for many years, the technology to detect polymorphisms has, until recently, been limited in resolution. Resolution in this context refers both to the ability to distinguish small changes (i.e. SNPs), and the ability to distinguish closely related strains from each other (i.e.. different isolates of the same strain). Note that it is sometimes desirable, depending on the forensic context, to use a method that can rapidly and inexpensively resolve isolates at the species level (i.e.. *B. subtilis* vs. *B. anthracis*), even if that method lacks the ability to resolve isolates. In contrast, a deeper investigation, without strict time constraints, might take advantage of slower, more expensive methods that have higher resolution.

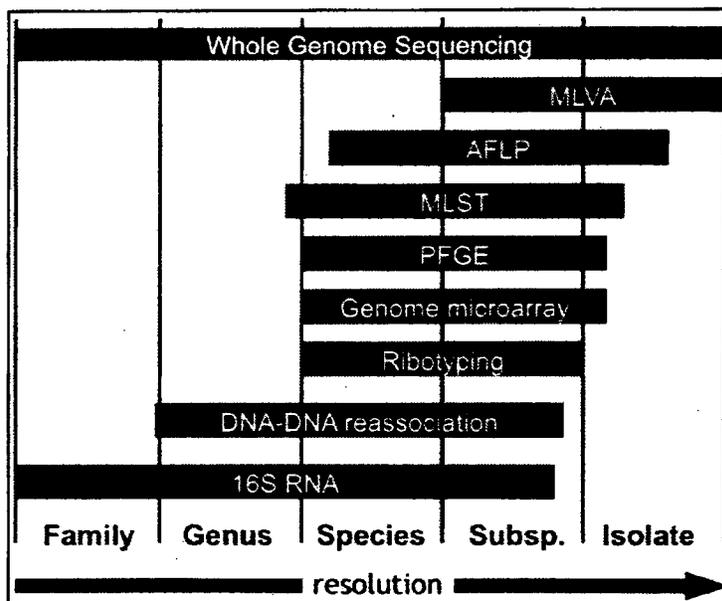


Figure 6. Resolution of polymorphism detection methods. PFGE, pulsed field gel electrophoresis; MLST, multi-locus sequence typing; AFLP, amplified fragment length polymorphism; MLVA, multi-locus VNTR analysis. Courtesy of Paul Keim.

Figure 6 compares methods of polymorphism detection, contrasting their abilities to resolve the relatedness of microbial organisms. Except for whole genome sequencing, the methods rely on assaying limited regions of the genome, often preselected for being highly polymorphic (Pereira et al., 2008). Because the complete sequence of the genome is the ultimate level of genetic resolution, whole genome sequencing is the only method that is informative at all levels of strain identity resolution.

Genome sequencing has been too slow, labor intensive, and expensive to use for routine microbial forensics investigations, but this is now changing with the development of new sequencing technologies, often referred to as next-generation sequencing (Mardis, 2008a, b). Figure 7 summarizes the differences between Sanger dideoxy sequencing, the most common method of DNA sequencing for the last 30 years, and several of the new technologies, listed by company responsible for their development. The new technologies differ most significantly from Sanger sequencing in that they bypass the need for cloning DNA fragments to be sequenced, relying instead on amplification of single molecules, or sequencing of single molecules without amplification. The new technologies are also massively parallel, yielding a massive increase in throughput relative to Sanger sequencing. Although currently expensive,

these next-generation sequencing technologies will, like all molecular biology methods, become less expensive so that it soon will be more cost effective to sequence an entire bacterial genome, rather than use older, less informative methods.

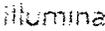
	 Sanger	 454 LIFE SCIENCES	 illumina	 Applied Biosystems
Isolate	Bacterial clone	PCR on beads	PCR colonies	PCR on beads
Parallelize	Capillary array	Microfab plate	Surface clusters	Magnetic surface
Sequencing method	Dideoxy chain termination	Pyrophosphate-induced fluorescence	Addition of cleavable fluorescent dNTPs	Stepwise ligation of cleavable fluorescent oligos
Read length (bp)	500–800	250	25–35	30–35
Reads / run	384	400,000	9,000,000	100,000,000
Throughput	300 kb / hour	300 Mb / day	600 Mb / day	3 Gb / 6 days

Figure 7. Comparison of Sanger sequencing with several next-generation sequencing technologies.

A benefit of next-generation sequencing is that the high throughput provides 30–100-fold coverage of a typical bacterial genome in a single sequencing run. Although each of the technologies is prone to systematic errors, this extremely deep sequencing eliminates most non-systematic sequencing errors simply by oversampling.

Application of next-generation sequencing technology to microbial forensics provides several benefits in addition to the increased throughput:

- 1) Next-generation sequencing technologies have the capability to sequence individual molecules, in contrast to Sanger sequencing which provides the “average” sequence of a large number of molecules. This capability can provide information about variation within a population because even variants that are present at a frequency of 0.05, for example, would potentially be sampled several times.
- 2) Whole genome sequencing can detect signatures of engineering that other methods cannot. There is increasing concern about the use of engineered organisms in a bioterrorism attack, a concern that is justified based on the ease with which many of the potential manipulations can be carried out. Signatures of engineering, such as selectable markers or polylinkers left in the genome, or altered phage or plasmid sequences, would

~~FOR OFFICIAL USE ONLY~~

be invisible to many of the methods that look at specific sites in the genome, but would readily be detected by whole genome sequencing.

- 3) Synthetic biology presents unpredictable difficulties for microbial forensics. The construction of an organism by the tools of synthetic biology might leave sequence evidence that could be detected by whole genome sequencing (such as sites used for stitching a genome together). However, it is likely that it will be possible to construct viral and bacterial genomes that have the exact sequence of historical pathogens (1918 influenza) or difficult-to-acquire pathogens (smallpox) with no signs of manipulation. In these cases it may be that the “perfection” of the sequence — the degree to which it matches the published sequence on which it was based — might itself be a signature.

Clearly, the trend is that whole genome sequencing will be the standard of analysis for both forensic and public health cases in which genetic identity of the isolate is important. The microbial forensics community should be at the forefront of this transition. This transition also creates a problem: vast amount of genome sequence will be generated by next generation sequencing technologies, and the sequence data are currently very different from previous sequence data. To maximally exploit these new data, research is needed to develop new bioinformatics tools to deal with short reads and deep coverage. An example is the development of a new version of the commonly used MUMmer alignment software, which runs on graphics processing units, and has improved performance on short-read data (Schatz et al., 2007).

### **3.3 Laboratory-based Experiments on Genetic Change vs. Natural Variation**

Much of microbial forensics can be accomplished using the naturally-occurring variation in microbial species to provide the polymorphisms needed for identification. In some cases, however, including in the scenario with which we were presented, one might want to know whether there are specific mutations that reproducibly occur during growth under particular conditions, such as in a patient, or in a culture grown according to a published protocol, or within a particular natural host. Knowledge of such reproducibly occurring variants might provide information regarding the previous history of a sample.

For most of the microbes of concern there is little information about the preferred trajectory for short-term genetic change. In most cases, the condition about which one would like information imposes a selection pressure on the microbial population, potentially resulting in genetic changes. In theory, it is possible to mimic a particular set of conditions in the laboratory, and to assay the changes that occur under those conditions. For example, given the above discussion of whole genome sequencing, one could use that technology to determine all changes that occur

~~FOR OFFICIAL USE ONLY~~

under a particular culture regime. Paul Keim described experiments from his laboratory in which serial passage experiments were carried out with a select agent bacterial species, such that approximately  $10^5$  total generations had occurred within 96 parallel lineages. Sequence analysis was performed on organisms obtained at the endpoint of each lineage to identify mutations common to multiple lineages, seeking to identify characteristic adaptive changes pertaining to the growth conditions that were used.

JASON was asked to consider whether this type of experiment should be expanded to other species and other growth conditions. Our assessment was that such experiments have limited utility and that they are unlikely to provide useful information except in very restricted circumstances. There are two major difficulties with such experiments. First, each experiment reflects a particular set of selection pressures, and the exact nature of those pressures often is difficult to determine. For example, in an experiment for which the goal is to reproduce conditions that might be used by an adversary growing a microbe for dissemination, there are variables that are likely to be important, but impossible to reproduce precisely, such as the source of water, the source of media components, the degree of aeration, and the consistency of incubation temperature. The second problem is that the sequence space available for evolution is vast, and all possible variants are unknowable in any experiment of reasonable scale. Advancing DNA sequencing technology, and the reduced cost of that technology, may change this assessment in detail, but the general limitations will continue to exist and are not limited by technology.

This negative assessment of laboratory evolution experiments is countered by the realization that nature is performing the same experiment, but on a much grander scale. We note that *Franciscella* isolates can easily be found in soil in many locations (Sellek et al., 2008), and from BioWatch air filtrates in at least two U.S. cities (Barns et al., 2005), indicating that it is widespread in the environment. Thus, collection and characterization of the relevant organisms from the wild will meet the goal of assessing permissive variation in microbial genomes. It will not meet the goal of knowing how a particular microbe changes its sequence under a particular set of growth conditions, but we conclude that this is not knowable by any practical experiment.

To fully exploit the natural variation in microbes of interest, we believe that a "Library of Congress" for microbial pathogens is needed. This library would consist of strains collected worldwide by methods that preserve sample properties, and capture all relevant data (e.g., geolocation, local environmental conditions). It should include laboratory isolates, natural isolates, and DNA sequence data. We were impressed with the efforts of the National

~~FOR OFFICIAL USE ONLY~~

Bioforensic Reference Collection along these lines. The NBRC was initiated in October 2005 to receive and store reference materials for forensic analyses. It currently has more than 30,000 samples of bacteria, viruses, and toxins, from both select and non-select agents, and is authorized to handle classified materials.

~~FOR OFFICIAL USE ONLY~~

#### 4 APPLICATION TO THE ATTACK SCENARIO

JASON was asked to consider the attack scenario described above, in which *F. tularensis* was disseminated at three locations, infected people at each location, and was found in a possible dissemination device. The scenario introduces several issues with respect to the ability to determine relatedness between isolates of the organism, the integrity of samples, and the ability to compare information from different types of samples. Here we focus on the questions posed in the scenario.

*Was a single preparation of material used for all of the attack locations, or were multiple preparations made?*

In addressing this question, we will assume for simplicity that there are samples available from each of the locations that represent the actual material released. The situation would be similar to asking whether the *B. anthracis* spores from the two temporally separate letter mailings in the 2001 anthrax case were from the same or different preparations. The first issue is genetic relatedness of the isolate strains. Depending on the amount of material available, it might be necessary to culture the isolates before applying genetic tests. As described above, we would propose that for such a case whole genome sequencing be used to assess relatedness of the various isolates. If the isolates from the three locations were derived from a single preparation of material, then they would be expected to be identical genetically. That is, all major and minor variants, or contaminants of other species, would be expected to be present in the same proportions. Note that because the material for such an attack would likely be produced by a large scale culture, there would be genetic variation within the population of cells (see Figure 5), and that this variation itself would provide a signature of relatedness, and that this signature could only be discerned fully by whole genome sequences. Of course, other evidence might also bear on whether the isolates came from the same preparation, such as the nature of any non-biological material in the preparation.

*Can the samples from patients be related between attack sites, and between patients and the dissemination device?*

To address this question, *F. tularensis* from patients at the three locations would be cultured and subjected to whole genome sequencing. We assume that growth in a human host is a selective environment in which genetic changes will occur, and therefore that the isolates might not be genetically identical, as defined in the example above. This lack of identity might be manifest both as the appearance of unique mutations in one population and not the others, and as a change in the ratios of variants when compared to the initial inoculum material. However, because the rate of mutation for any individual base pair is very low, it is likely that the set of SNPs defining the initial inoculum isolate would still be present in the patient isolates. The presence of these SNPs might be definitive as to the relationship of the patient strains, depending on the prevalence

~~FOR OFFICIAL USE ONLY~~

of those SNPs in environmental populations. It would be helpful for this purpose to have better knowledge about the genetic variation within natural populations of *F. tularensis*, as suggested above.

*Can samples of F. tularensis from domestic laboratories, or IC sources be definitively linked to the attack material, or excluded, by genetic similarity?*

This question strikes at the heart of the issue for forensics: the attempt to determine whether samples in hand from other collection efforts match those used in an attack. The question specifically asks whether the samples can be “definitively linked to the attack material”. The limitations on genetic relatedness in asexually reproducing organisms were discussed above. JASON concludes that these limitations are such that *it is never possible to definitively link a sample to an attack based on genetic evidence alone*. It is possible to conclude that a sample is a close (or identical) match to the attack sample, and therefore that the sample might have been the source of the attack material. A meaningful quantitative assessment of any identified genetic relatedness is confounded by the many unknowable variables that might affect relatedness. For example, a 100% match of whole genome sequence data from sample and attack material would not definitively link that particular sample to the attack because there could have been a very large number of such samples prepared at once, then disseminated in location and kept indefinitely in freezers. Similarly, a less than 100% match does not rule out a sample as having been the root source material for an attack, as there may have been many generations of growth under unknown conditions separating the collected sample from the original attack material. Of course, the genetic analysis would be done in the context of a broader forensics investigation, and a measure of similarity would likely be useful in that context.

*If genetic relatedness is inherently difficult to use to make definitive links between samples, can it be used more effectively to exclude samples from having been associated with an attack?*

Here the estimates of mutation rate, as derived from studies of neutral mutations, are directly relevant (Lenski and Keim, 2005). As discussed above, a strain that is derived from a parent strain is likely to retain neutral mutations found in the parent because the mutation rate for any particular base pair is very low. Thus, loss of several unique markers that identify the parent strain would typically be very unlikely, and might, in a quantitatively defensible way (Lenski and Keim, 2005), be used to exclude a collected strain from being the source of the attack material. Note that this assumes no bioengineering of the strain by the perpetrators of the attack, because knowledge of the genetic variants identifying particular strains would allow these markers to be changed, yielding a false exclusion.

~~FOR OFFICIAL USE ONLY~~

*Is geolocation based on genetic sequence analysis feasible and could it be useful for attribution?*

Although there are some small-scale success stories in geolocation within confined areas (P. Keim, unpublished results), we conclude that the worldwide distribution of microbes by wind, planes, ships, and people (intentional or unintentional), would make geolocation difficult to use in a forensics investigation. However, it is clear that there are geographically restricted genotypes for many microbes (see (Blackburn et al., 2007) for a *B. anthracis* example), and we suggest the potential for genetic geolocation be revisited when there are more data from current and future metagenomics projects. These projects aim to sample broadly the microbial diversity in many locations, and have the potential to aid in an understanding of the geographical distribution of variants that could be useful for geolocation, with the caveats discussed above. As an example of this approach, Venter and colleagues have collected and sequenced organisms from a large swath of the Atlantic and Pacific (Rusch et al., 2007; Williamson et al., 2008; Yooseph et al., 2007). Although the sites for collection were not chosen with a particular scientific question in mind, different locations in the ocean were found, perhaps not surprisingly, to have different populations of microbes. It remains to be seen whether these differences would be reproducible.

As a side note, we anticipate that metagenomics projects will reveal sites at which it would be possible to recover threat organisms, even particular variants of such organisms, from the environment. The data from these projects are usually publicly accessible, and can be searched for sequences corresponding to any organism. As these projects become more commonplace, and collection points have geographical tags associated with them, it may become a relatively simple matter to isolate a strain of a select agent from a natural habitat identified by metagenomic sequencing. A related concern is that simple searches of the metagenomic sequence information are likely to reveal variants of known pathogenicity genes. These variants could be easily synthesized, without recovering the organism from which they are derived, and might differ from the characterized genes in virulence in human infection, or in recognition by existing vaccines.

Lastly, although we were tasked with assessing the use of genetic methods in microbial forensics, it is clear that non-genetic signatures might also be of use in a forensics investigation. There has been some investigation of stable isotope ratios as a tool for geolocation (Kreuzer-Martin et al., 2004a, b; Kreuzer-Martin et al., 2005), although we have seen little evidence that this will be broadly applicable. However, it is possible that the power of next-generation sequencing will provide a "metagenomics signature" for cases in which either the attack material is available, or in which a device is recovered, which could be examined for contaminating

~~FOR OFFICIAL USE ONLY~~

microorganisms. The profile of the low-level microbial contaminants in such material, detectable with next-generation sequencing, might help to identify the source.

~~FOR OFFICIAL USE ONLY~~

## 5 CONCLUSIONS AND RECOMMENDATIONS

The primary conclusions of the study are that:

- 1) Microbial forensics is a powerful tool, but is most useful as a complement to field work, human intelligence, evidence gathering, and other traditional forms of forensics.
- 2) Whole-genome DNA sequencing provides the ultimate level of genetic resolution for forensics and will replace the current forensic methods of analyzing DNA polymorphisms. It is also revolutionizing metagenomics. The microbial forensics community should be at the forefront of this transition.
- 3) A quantitative assessment of the meaningfulness of a “match” for forensics is difficult, because of the asexual growth of bacteria, lack of information about population structure, mutation rate, and number of generations that separate two samples, and the unknown effects of human intervention.
- 4) Quality of relatedness information (phylogenetic trees) is limited by the quantity, quality, and diversity of the data collected.
- 5) Although DNA-based methods of strain comparison are the most powerful, methods that exploit other signatures will likely be important forensically.
- 6) Improving microbial forensics capabilities also improves capabilities for analyzing natural outbreaks, and *vice versa*.

The chief recommendations of the study are the following:

- 1) For all relevant pathogens, the National Biodefense Forensic Analysis Center (NBFAC) should sample the diversity of both natural and laboratory strains, and sequence at low fidelity to get a sense of sequence variation. A few representative strains should be chosen for high-fidelity sequencing to establish “gold standard” sequences for each pathogen. High-fidelity sequence of a related outgroup should also be acquired (*B. cereus* for *B. anthracis*, for example).
- 2) The broad group of government agencies with an interest, including NBFAC, IC, FBI, CDC, and USDA, should develop and promote best practices for microbial sample collection that best preserve the genetic and non-genetic signatures of interest.
- 3) NBFAC, with its bioinformatics partners at LLNL, should exploit data from emerging metagenomic projects to expand knowledge of natural diversity. Bioinformatics tools better able to deal with this growing sequence database are needed. Some of these are

~~FOR OFFICIAL USE ONLY~~

currently being developed by the academic community, others by the commercial vendors of sequencing technology, but some may need to be developed in-house.

- 4) The strain collection efforts of NBFAC and the National Bioforensic Reference Collection should be expanded, with greater support, and coupled with the development of a relational database that includes genetic, geographic, and other contextual information. Sequences themselves, in addition to organisms, should be considered part of this database.

JASON also identified two longer-term basic research goals that will impinge upon microbial forensics capabilities:

- 1) Develop capabilities for novel microbial forensics methods. Examples of possible directions to develop new signatures useful for forensics are a) transcriptome analysis - the profile of the RNA molecules that are expressed from the genome, which is known to vary with growth condition; b) epigenetic modifications — covalent modifications of the genomic DNA that control gene expression under different conditions; and c) metabolomics — the profile of the small organic molecules present in cells.
- 2) Develop the “pathogenome” (the molecular and genetic mechanisms of pathogenicity) as an organizing concept for microbial pathogens, complementing “species” organization. The rapid horizontal transfer of genetic elements amongst bacterial species has altered the view of the nature of species in the prokaryotic kingdom. Indeed, it is possible for relatively harmless bacterial species, such as *B. cereus*, to acquire pathogenic properties from a pathogenic species, *B. anthracis*, and to cause an anthrax-like disease in humans (Avashia et al., 2007; Hoffmaster et al., 2006; Hoffmaster et al., 2004). This suggests that, rather than the organism, we should consider the pathogenic mechanism and the DNA (and RNA) molecules encoding it as the relevant agent. This organization also helps to free one from lists such as the Select Agent list that single out one organism that bears a particular pathogenic mechanism, but not others that might also be able to.

## 6 REFERENCES

1. Avashia, S.B., Riggins, W.S., Lindley, C., Hoffmaster, A., Drumgoole, R., Nekomoto, T., Jackson, P.J., Hill, K.K., Williams, K., Lehman, L., *et al.* (2007). Fatal pneumonia among metalworkers due to inhalation exposure to *Bacillus cereus* Containing *Bacillus anthracis* toxin genes. *Clin Infect Dis* 44, 414-416.
2. Baechtel, F.S., Monson, K.L., Forsen, G.E., Budowle, B., and Kearney, J.J. (1991). Tracking the violent criminal offender through DNA typing profiles--a national database system concept. *EXS* 58, 356-360.
3. Barns, S.M., Grow, C.C., Okinaka, R.T., Keim, P., and Kuske, C.R. (2005). Detection of diverse new *Francisella*-like bacteria in environmental samples. *Appl Environ Microbiol* 71, 5494-5500.
4. Beckstrom-Sternberg, S.M., Auerbach, R.K., Godbole, S., Pearson, J.V., Beckstrom-Sternberg, J.S., Deng, Z., Munk, C., Kubota, K., Zhou, Y., Bruce, D., *et al.* (2007). Complete genomic characterization of a pathogenic A.II strain of *Francisella tularensis* subspecies *tularensis*. *PLoS ONE* 2, e947.
5. Blackburn, J.K., McNyset, K.M., Curtis, A., and Hugh-Jones, M.E. (2007). Modeling the geographic distribution of *Bacillus anthracis*, the causative agent of anthrax disease, for the contiguous United States using predictive ecological [corrected] niche modeling. *Am J Trop Med Hyg* 77, 1103-1110.
6. Brotcke, A., Weiss, D.S., Kim, C.C., Chain, P., Malfatti, S., Garcia, E., and Monack, D.M. (2006). Identification of *MglA*-regulated genes reveals novel virulence factors in *Francisella tularensis*. *Infect Immun* 74, 6642-6655.
7. Denamur, E., and Matic, I. (2006). Evolution of mutation rates in bacteria. *Mol Microbiol* 60, 820-827.
8. Hoffmaster, A.R., Hill, K.K., Gee, J.E., Marston, C.K., De, B.K., Popovic, T., Sue, D., Wilkins, P.P., Avashia, S.B., Drumgoole, R., *et al.* (2006). Characterization of *Bacillus cereus* isolates associated with fatal pneumonias: strains are closely related to *Bacillus anthracis* and harbor *B. anthracis* virulence genes. *J Clin Microbiol* 44, 3352-3360.
9. Hoffmaster, A.R., Ravel, J., Rasko, D.A., Chapman, G.D., Chute, M.D., Marston, C.K., De, B.K., Sacchi, C.T., Fitzgerald, C., Mayer, L.W., *et al.* (2004). Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc Natl Acad Sci U S A* 101, 8449-8454.

~~FOR OFFICIAL USE ONLY~~

10. Kreuzer-Martin, H.W., Chesson, L.A., Lott, M.J., Dorigan, J.V., and Ehleringer, J.R. (2004a). Stable isotope ratios as a tool in microbial forensics--Part 1. Microbial isotopic composition as a function of growth medium. *J Forensic Sci* 49, 954-960.
11. Kreuzer-Martin, H.W., Chesson, L.A., Lott, M.J., Dorigan, J.V., and Ehleringer, J.R. (2004b). Stable isotope ratios as a tool in microbial forensics--Part 2. Isotopic variation among different growth media as a tool for sourcing origins of bacterial cells or spores. *J Forensic Sci* 49, 961-967.
12. Kreuzer-Martin, H.W., Chesson, L.A., Lott, M.J., and Ehleringer, J.R. (2005). Stable isotope ratios as a tool in microbial forensics--part 3. Effect of culturing on agar-containing growth media. *J Forensic Sci* 50, 1372-1379.
13. Larsson, P., Oyston, P.C., Chain, P., Chu, M.C., Duffield, M., Fuxelius, H.H., Garcia, E., Halltorp, G., Johansson, D., Isherwood, K.E., *et al.* (2005). The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nat Genet* 37, 153-159.
14. Lauriano, C.M., Barker, J.R., Yoon, S.S., Nano, F.E., Arulanandam, B.P., Hassett, D.J., and Klose, K.E. (2004). MglA regulates transcription of virulence factors necessary for *Francisella tularensis* intraamoebae and intramacrophage survival. *Proc Natl Acad Sci U S A* 101, 4246-4249.
15. Lenski, R.E., and Keim, P. (2005). Population Genetics of Bacteria in a Forensic Context. In *Microbial Forensics*, R.G. Breeze, B. Budowle, and S.E. Schutzer, eds. (Amsterdam, Elsevier).
16. Mardis, E.R. (2008a). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9, 387-402.
17. Mardis, E.R. (2008b). The impact of next-generation sequencing technology on genetics. *Trends Genet* 24, 133-141.
18. Pereira, F., Carneiro, J., and Amorim, A. (2008). Identification of species with DNA-based technology: current progress and challenges. *Recent Pat DNA Gene Seq* 2, 187-199.
19. Petrosino, J.F., Xiang, Q., Karpathy, S.E., Jiang, H., Yerrapragada, S., Liu, Y., Gioia, J., Hemphill, L., Gonzalez, A., Raghavan, T.M., *et al.* (2006). Chromosome rearrangement and diversification of *Francisella tularensis* revealed by the type B (OSU18) genome sequence. *J Bacteriol* 188, 6977-6985.

~~FOR OFFICIAL USE ONLY~~

20. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5, e77.
21. Schatz, M.C., Trapnell, C., Delcher, A.L., and Varshney, A. (2007). High-throughput sequence alignment using Graphics Processing Units. *BMC Bioinformatics* 8, 474.
22. Sellek, R., Jimenez, O., Aizpurua, C., Fernandez-Frutos, B., De Leon, P., Camacho, M., Fernandez-Moreira, D., Ybarra, C., and Carlos Cabria, J. (2008). Recovery of *Francisella tularensis* from soil samples by filtration and detection by real-time PCR and cELISA. *J Environ Monit* 10, 362-369.
23. Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., *et al.* (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3, e1456.
24. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 5, e16.