REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
15-07-2014	Final Report	25-Sep-2012 - 24-Mar-2014

4. TITLE AND SUBTITLE
Research Area 7.4:Identifying a Path Towards Rapid Discrimination of Infection Disease Outbreaks

5a. CONTRACT NUMBER
W911NF-12-1-0599

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS
Stephen Turner, Carol Gilchrist, Margaret Riley, William Petri, Erik Hewlett

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Virginia
P. O. Box 400195

Charlottesville, VA 22904 -4195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
63109-CH.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The low cost and relative ease of obtaining, producing, and disseminating pathogenic organisms or biological toxins in an act of bioterrorism is a signi?cant concern in the United States and other parts of the world. The United States Government began a new civilian biodefense program as early as 1996, motivated by a combination of (1) high-pro?le terrorist events in the U.S., (2) the extent of chemical and biological warfare program development in Iraq and the former Soviet Union, and (3) both real and ?ctional accounts of biological threats to the American population. Using next-generation sequencing and bioinformatics technologies to sequence and analyze the whole

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Erik Hewlett
UU	UU	UU	UU		19b. TELEPHONE NUMBER 434-924-5945

Report Title

Research Area 7.4:Identifying a Path Towards Rapid Discrimination of Infection Disease Outbreaks

ABSTRACT

The low cost and relative ease of obtaining, producing, and disseminating pathogenic organisms or biological toxins in an act of bioterrorism is a signi?cant concern in the United States and other parts of the world. The United States Government began a new civilian biodefense program as early as 1996, motivated by a combination of (1) high-pro?le terrorist events in the U.S., (2) the extent of chemical and biological warfare program development in Iraq and the former Soviet Union, and (3) both real and ?ctional accounts of biological threats to the American population. Using next-generation sequencing and bioinformatics technologies to sequence and analyze the whole genomes of emerging pathogens is a capability that can provide genetic analysis of pathogens in a microbial forensics investigation with the highest-possible resolution, and can assist in discriminating between natural, accidental, and deliberate causes for infectious disease outbreaks. Here we review NGS and bioinformatics technologies, giving an account of current, near, and long-term capabilities and limitations. Through a series of case studies, we show how NGS and bioinformatics technology can assist in determining whether an infectious disease outbreak is due to accidental, deliberate, or intentional causes and provide recommendations on a path forward to future deployable capability.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received	Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Post Doctorates

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Faculty Supported

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: .....

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):.....

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:.....

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense .....

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: .....

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

No changes were suggested or requested by the Sponsor, after the report was attached and submitted to the 09/27/2013 Interim ARO Progess Report. This report is being sent to "Clinical Microbiology" for final reviews.

Technology Transfer

Final Report for: Research Area 7.4:Identifying a Path Towards Rapid Discrimination of Infection Disease Outbreaks

Harnessing Next-Generation Sequencing Capabilities for Microbial Forensics

Stephen D. Turner, Ph.D.[*]
Carol Gilchrist, Ph.D.[†]
Margaret Riley, J.D.[‡]
William A. Petri Jr., M.D.[§]
Erik Hewlett, M.D.[¶]

September, 2013
Charlottesville, Virginia

Contents

[*]Department of Public Health Sciences, University of Virginia School of Medicine.
†Department of Medicine, University of Virginia School of Medicine.
‡Department of Public Health Sciences, School of Medicine, Batten School of Leadership & Public Policy, and School of Law, University of Virginia.
[§]Departments of Medicine, Microbiology, and Pathology, University of Virginia School of Medicine.
[¶]Departments of Medicine and Microbiology, University of Virginia School of Medicine.

Summary

The low cost and relative ease of obtaining, producing, and disseminating pathogenic organisms or biological toxins in an act of bioterrorism is a significant concern in the United States and other parts of the world. The United States Government began a new civilian biodefense program as early as 1996, motivated by a combination of (1) high-profile terrorist events in the U.S., (2) the extent of chemical and biological warfare program development in Iraq and the former Soviet Union, and (3) both real and fictional accounts of biological threats to the American population (Khan, Levitt, & Sage, 2000; Treadwell, Koo, Kuker, & Khan, 2003). Using next-generation sequencing (NGS) and bioinformatics technologies to sequence and analyze the whole genomes of emerging pathogens is a capability that can provide genetic analysis of pathogens in a microbial forensics investigation with the highest-possible resolution, and can assist in discriminating between natural, accidental, and deliberate causes for infectious disease outbreaks. Here we review NGS and bioinformatics technologies, giving an account of current, near, and long-term capabilities and limitations. Then, through a series of case studies, we show how NGS and bioinformatics technology can assist in determining whether an infectious disease outbreak is due to accidental, deliberate, or intentional causes and provide recommendations on a path forward to future deployable capability.

Report goals:

- To review the state-of-the-art genome-sequencing technology and bioinformatics techniques necessary for analyzing high-throughput genomic data in microbial forensic applications.
- To analyze recent case studies from the public health literature, in which next-generation, whole-genome sequencing (NGWGS) is used in transmission network reconstruction and source attribution for emerging pathogens.
- To identify issues to consider when using genome sequencing data to investigate whether an infectious disease outbreak was initiated deliberately, accidentally or naturally.
- To recognize ways to enhance the value of genome sequencing to microbial forensics.
- To identify gaps and outline a path forward to a field-deployable capability harnessing DNA sequencing technology for microbial forensics applications.

Acronyms and definitions:

- **Assembly (genome assembly)**: the process of taking a large number of short DNA sequence *reads* and piecing them back together to create a contiguous representation of the original DNA sequence from which the reads originated.
- **bp**: Base-pairs. I.e., nucleotides (A, C, G, or T) that make up a DNA sequence. The length of DNA sequence reads, assemblies, genomes, or sequencer output is referred to in terms of the length in bp, often prefixed by standard metric units. E.g., Mb=megabases=1 million bp; Gb=gigabases=1 billion bp; etc.
- **MRSA**: Methicillin-resistant *Staphylococcus aureus*. MRSA is a bacterium responsible for several difficult-to-treat infections in humans.
- **Metagenomics**: The study of metagenomes, or DNA collected directly from the environment. Most microbes are unculturable. Sequencing DNA collected directly from the environment, paired with powerful bioinformatics analysis represents another forensic signature that can be exploited for microbial forensics attribution.
- **NGS**: Next-Generation Sequencing. Newer methods of DNA sequencing that usually rely on sequencing-by-synthesis rather than dye-terminator (a.k.a. “Sanger”) methodology. The cost of DNA sequencing has decreased by several orders of magnitude since the advent and continual improvement of NGS technologies. NGS employs massively parallel high-throughput generation of short (50-250 bp) sequencing *reads*. Technologies, such as SMRT sequencing offered by Pacific Biosciences, or the technology being developed by Oxford Nanopore, are capable of sequencing very long reads at the expense of throughput, colloquially referred to as “3rd-gen sequencing,” can be included in NGS technologies.
- **Read**: a sequence *read* is a short sequence of nucleotides (typically 50-250 bp for an Illumina instrument), usually connected to each nucleotide’s respective base-call quality.
- **WGS**: Whole-Genome Sequencing. WGS refers to using sequencing technology (usually NGS) to sequence the entire genome of an organism to sufficiently high depth-of-coverage so as to enable genome assembly.

Acknowledgements

This work was supported by a grant from the Department of Defense, Army Research Office to Dr. E. Hewlett.

Task 1: Critical assessment of next-generation sequencing and bioinformatics technology: capabilities and limitations as applied to microbial forensics applications

Next-generation sequencing (NGS) technology

The advent and continual development of instrumentation for inexpensive ultra high-throughput next-generation DNA sequencing (NGS) has transformed microbial genome sequencing from a multimillion-dollar, team-science enterprise into a routine exercise in molecular biology. Small "benchtop" sequencers that have been on the market for the last three years are capable of producing several gigabases of sequence data per run, and have more than sufficient capacity for sequencing a single microbial genome to a high depth of coverage.

An in-depth discussion of the technological approaches and scientific details of NGS chemistry is beyond the scope of this report, and are reviewed in depth elsewhere (Metzker, 2010). Briefly, the 454 GS Junior is a smaller, lower-throughput version of the 454 GS FLX instrument, utilizing emulsion PCR and pyrosequencing technologies. The Ion Torrent PGM also uses emulsion PCR, but uses sequencing-by-synthesis and semiconductor technology to detect protons released as nucleotides are incorporated during synthesis. Illumina currently dominates the benchtop sequencer market share with its sequencing-by-synthesis approach using fluorescently labeled reversible-terminator nucleotides on amplified DNA fragments immobilized on a glass slide. Illumina's MiSeq instrument is a smaller, faster, lower-throughput version of its flagship HiSeq 2500 instrument.

While not benchtop in size, the Pacific Biosciences (PacBio) RS instrument is the first of its kind to utilize single molecule real time (SMRT) sequencing, where each base is detected in real time as a polymerase adds fluorescently tagged nucleotides along individual DNA template molecules. SMRT sequencing is distinct from the other technologies mentioned here due to its very long read lengths. The read lengths of the current PacBio RS instrument can reach over 20,000 bp, with an average of about 3,000 bp, which is 30-200 times longer than read lengths from other NGS instruments, and four times longer than the original release of the instrument in 2011 (Roberts, Carneiro, & Schatz, 2013), and the PacBio RS II instrument promises even longer read lengths (http://pacificbiosciences.com/). Furthermore, examining the polymerase kinetics of SMRT sequencing allows for direct detection of methylated DNA bases (Flusberg et al., 2010). This intrinsic capability of the SMRT sequencing workflow does not affect the primary sequence determination, while yielding yet another forensic signature that is not captured with standard protocols on other NGS instruments. Finally, while the single-pass error rate of SMRT sequencing is exceptionally high compared to other NGS platforms, the accuracy of a genome sequence determined by consensus of overlapping or circular fragments is >99.999%. (C.-S. Chin et al., 2013; Koren et al., 2012). While SMRT throughput is much lower than that of other NGS technologies, the longer read lengths and direct detection of DNA methylation makes it ideal for certain forensic sequencing applications. These advantages will be discussed further in the bioinformatics section below.

Several in-depth comparisons of benchtop sequencing instruments have been published over the last year, specifically comparing the Roche 454 GS Junior, Life Technologies Ion Torrent PGM, and the Illumina MiSeq (Jünemann et al., 2013; Loman et al., 2012). In the 14 months that passed between the Loman comparison (Loman et al., 2012) and the Junemann analysis (Jünemann et al., 2013), throughput increased approximately 5-fold and read lengths quadrupled (Pallen, 2013) at a faster-than-Moore's-Law pace (Wetterstrand, 2013). These comparisons showed that the 454 GS Junior's throughput per run was too low to faithfully assemble typical bacterial genomes, and that the MiSeq is best with respect to maximizing throughput and minimizing consensus errors. Another recently

published paper compared whole-genome sequencing data on three different microbial pathogens from the MiSeq, Ion Torrent, and the Pacific Biosciences RS instruments (Quail et al., 2012), reporting on coverage distribution, bias, GC distribution, variant detection, and accuracy, for each instrument on each of the three genomes. This study reported that sequences generated by the Ion Torrent, MiSeq and Pacific Biosciences technologies all displayed near perfect coverage behavior on GC-rich, neutral and moderately AT-rich genomes. However, a profound bias was observed upon sequencing the extremely AT-rich genome of *Plasmodium falciparum* on the Ion Torrent PGM, resulting in no coverage for approximately 30% of the genome. This study also reported on the ability to call variants from each platform and found that slightly more variants could be identified from Ion Torrent data, compared to MiSeq data, but at the expense of a higher false positive rate. Finally, this comparison highlighted the utility of PacBio's long reads for *de novo* assembly, at the expense of a lower throughput and single-pass accuracy, making the technology a poor selection for resequencing and finding rare variants.

A new sequencing platform, rumored to be completely revolutionary, is in development by Oxford Nanopore. Here, a single DNA molecule is passed through a nanopore set in an electrically resistant membrane bilayer, and fluctuations in the current across the membrane can be used to identify the molecule (nucleotide) in question (Eisenstein, 2012). The company projects that bases will be measured starting at 100-200 bases per second, up to 1000 bases per second in the future, with read lengths up to tens of thousands of bases long (Didelot, Bowden, Wilson, Peto, & Crook, 2012). The system is designed to work with native DNA (not specially-prepared libraries), so the company purports that very crude samples (e.g. mud, blood, etc.) with low DNA concentrations can be loaded directly onto the instrument. The company has announced two different platforms: the larger scalable GridION system, capable of generating 2 gigabases of data per hour, and the single-use USB-connected MinION system, capable of generating 150 megabases of sequence per hour. If accuracy of these systems can match that currently achieved by state-of-the-art next-generation sequencing technology, long reads will allow bacterial genomes to be completed within minutes. Nanopore sequencing technology has been under development since 1995 (Kasianowicz, Brandin, Branton, & Deamer, 1996), but if the potential of this technology is realized, new bioinformatics algorithms capable of leveraging real-time streaming long reads must be developed in order to fully exploit this new technology, and should be an area of focus for future research and development.

The "holy grail" of DNA sequencing technologies would have the following features: (1) high accuracy, (2) high throughput, (3) long read lengths, and (4) rapid turnaround time. Current technologies only allow for two or three of those four qualities. Operational constraints for some microbial forensics applications impose further requirements such as having a small footprint, low power requirements, and the ability to be ruggedized. The National Human Genome Research Institute (NHGRI) at the National Institutes of Health regularly updates an analysis of the cost of DNA sequencing (Wetterstrand, 2013). This analysis presented at http://genome.gov/sequencingcosts/ compares the cost of sequencing to Moore's Law (the observation that computing power doubles every two years), showing that sequencing throughput has increased much faster than Moore's Law. That is, until very recently – in the third quarter of 2012 the cost of sequencing actually *increased* by 12% for the first time since records began. Doubtlessly this trend is only a temporary change in the inevitable fall of sequencing costs, but it has several implications for the utility of NGS in microbial forensics as outlined in (N. Hall, 2013). This temporary leveling out of sequencing costs means that what is possible next year will not be drastically or qualitatively different than what is possible now, and that no new doors will be opened by substantially lower sequencing costs. However, this period of stability will ultimately be beneficial for the application of NGS to microbial forensics applications, because bioinformaticians will be able to hone the methods we are currently developing and address the fundamental challenges we currently face, without the overbearing threat of a disruptive new technology suddenly rendering everything we are doing completely obsolete. These bioinformatics challenges and opportunities are discussed in further detail below.

Bioinformatics

Microbial forensics and diagnostic bacteriology have historically been dependent on isolating and culturing a single viable organism, and the principles behind microbial identification have changed very little over the last fifty years (Didelot et al., 2012). These methods include selective culturing, serotyping, biochemical testing, Gram staining, mass spectrometry, coupled with expert evaluation to make an assessment about species and pathogenicity. This process is costly, labor-intensive, and can take months from sample collection to final attribution. Further, these methods rely on the availability of a pure culture; the vast majority of microorganisms are unculturable (Rappé & Giovannoni, 2003; Wooley, Godzik, & Friedberg, 2010), and many microbial forensics samples will consist of a mixture of microorganisms (see the Shotgun Metagenomics section below).

Ideally, all of the information needed for an effective attribution and appropriate response would be attained in a single step. In principle, a microorganism's genome sequence contains all this information, and rapid, inexpensive NGS (see above) has the potential to replace many of the above-mentioned labor-intensive procedures currently used for attribution in microbial forensics. As discussed above, NGS is a highly disruptive new technology that has the potential to revolutionize microbial forensics, replacing many of the cumbersome and laborious conventional bacteriological methods with higher-resolution genome-based information. However, the bottleneck in data-to-knowledge is no longer at the data generation phase – there are substantial bioinformatic challenges to be met, and success of a microbial forensics program will depend on development of the tools and reference databases necessary to extract and interpret this information rapidly and correctly.

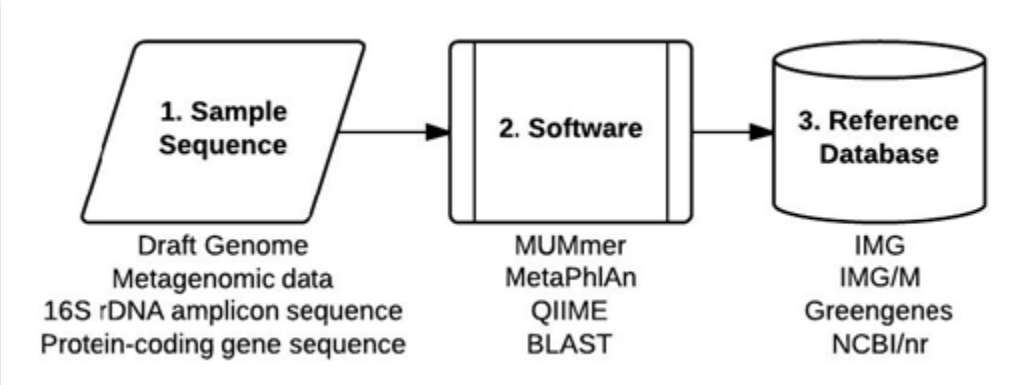

Figure 1: Conceptual representation of a Microbial Forensics Bioinformatics Pipeline. There are three components necessary for microbial forensic analysis with NGS data. The first is a DNA sequence of the sample (1), which could be a completed draft genome assembly, shotgun metagenomic sequences, or amplicon sequences such as 16S rDNA or protein coding genes. An algorithm/software implementation (2) is the necessary link between the sample sequence data (1) and a comprehensive reference database (3) consisting of all available draft and complete genomes, metagenomes, 16S rRNA gene sequences, or other protein-coding gene sequences.

Microbial forensic analysis with NGS data requires three components (Figure 1): 1) sequence information for the sample of interest, 2) a comprehensive reference database for comparison (single genomes, metagenomes, 16S rDNA sequences, geographically-indexed databases, etc.), and 3) an algorithm and software implementation to probabilistically link sequences from the sample (Figure 1: #1) to the reference database (Figure 1: #3).

In this section we discuss the challenges, state-of-the-art solutions, and existing analytical gaps for microbial forensics using NGS data. We make an important distinction between (A) identification and attribution of a single organism from a pure culture, and (B) shotgun sequencing of an environmental metagenomic sample containing a plurality of distinct microorganisms.

Pure culture / single isolate sample

While not trivial, bioinformatics analysis of sequencing data from pure cultures of a single microorganism is considerably easier than with contaminated, mixed, or metagenomic samples. Modern benchtop sequencers like the MiSeq do not directly produce a contiguous, complete sequence of an organism's genome. Instead, they produce many short reads, typically 100 to 250 bp in length, randomly sampled many times from different parts of the genome. Genome assembly is the process of combining short reads into larger contiguous sequences called contigs (Figure 2).

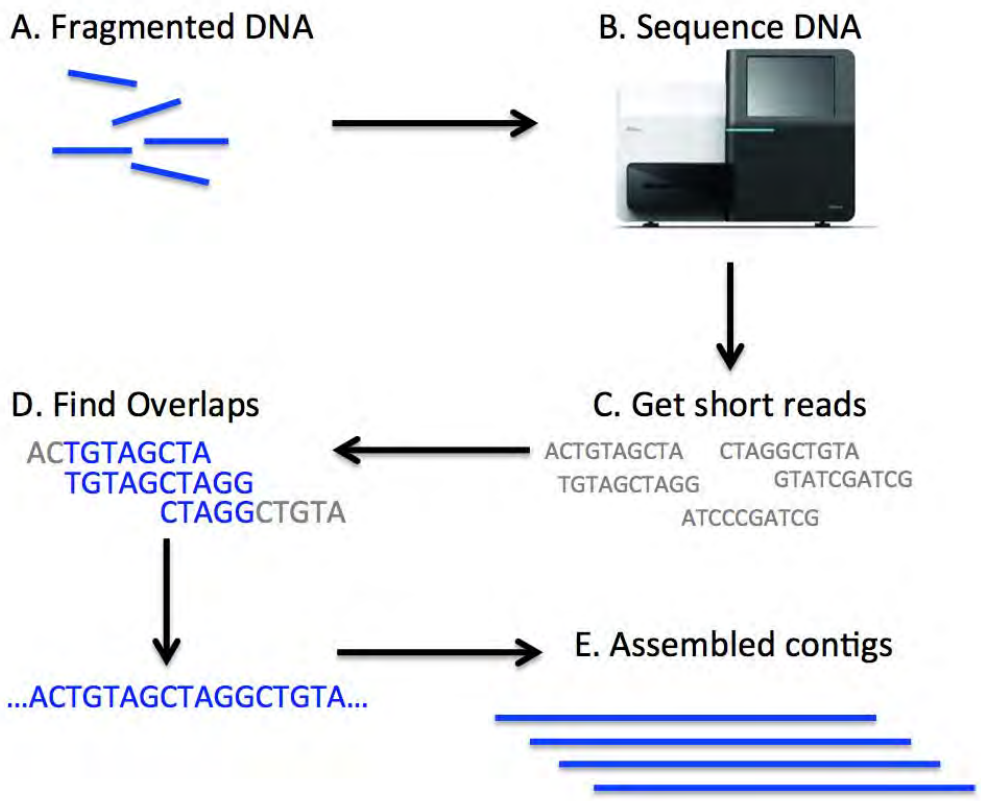

Figure 2: Raw reads vs. assembled genomes. Fragmented DNA (A) is sequenced to generate millions of 100-250bp short reads (C). Cleaning the data and assembling reads based on overlapping sequence results in long *contigs* (E), which are often >100,000bp in length. The raw output from the sequencer consists of short reads (C), forensic analysis can be more in-depth using assembled contigs (E).

A large number of both commercial and free/open-source high-performance software tools exist for genome assembly (Gnerre et al., 2011; R. Li et al., 2010; Y. Peng, Leung, Yiu, & Chin, 2012; Simpson et al., 2009; Zerbino & Birney, 2008). However, recent genome assembly "competitions" such as GAGE-B (Magoc et al., 2013) and the Assemblathons (K. R. Bradnam et al., 2013; Earl et al., 2011) have consistently observed that n*o* singl*e* assembl*er* outperform*s* a*ll* othe*r* assembler*s* i*n* a*ll* assembl*y* metric*s* fo*r* a*ll* organisms. Specifically, both of the competitions mentioned above found that (1) on a single organism, different assemblers ranked completely differently depending on which metric was used to judge the quality of the assembly; and (2) for the same metric, different assemblers ranked differently with different organisms. A detailed review of genome assembly can be found elsewhere (Baker, 2012).

After sequence reads have been assembled, standard sequence similarity tools can be used to compare assembled genomes to databases of both finished and draft microbial genomes (Delcher,

Salzberg, & Phillippy, 2003), which are available from sources such as NCBI and IMG (Markowitz et al., 2012). Standard tools are available for phylogenetic analysis, which is helpful for quantitatively and visually assessing evolutionary relationships, as was recently done during the 2011 *E. coli* O104:H4 outbreak in Germany (Grad et al., 2012; Mellmann et al., 2011). Basic Local Alignment Search Tool (BLAST) (Altschul, Gish, Miller, Myers, & Lipman, 1990) has been widely used for decades to perform similarity searches with sequence data. Even when a draft genome assembly is not available, BLAST searching from an amplified region of interest from a pure culture of a single organism against a reference database can also be used for putative identification of a sample species. BLAST searching against databases of virulence genes and genetic engineering vectors can determine the presence/absence of these factors. In these cases (and as is often the case with metagenomics, discussed below), the completeness of the reference database is often what limits the resolution one can achieve with microbial forensic analysis using NGS.

While the *de novo* assembly of genomes remains challenging, the volume of sequence data produced by NGS technology is usually high enough that several relatively small bacterial genomes can be sequenced in one run without sacrificing coverage depth. Here, an index barcode is used in the initial sample preparation step to 'tag' the fragments with a short sequence unique to the sample. This barcode can then be used to bin the sample reads according to the source (demultiplex) prior to downstream analysis. The number of samples that can be multiplexed depends on the target genome size and the depth of desired sequence coverage. For example, the Illumina MiSeq is capable of generating approximately 5 Gbp using a 2x150 paired-end run (about 24 hours sequencing run time). With 5 Gbp of sequence data, 10 samples can be multiplexed on a single run assuming desired $100X$ coverage of a 5 Mb bacterial genome (5×10^6 bp genome $\times$ $100X$ coverage $\times$ 10 samples = 5×10^9 bp MiSeq output). As the throughput of instruments like the MiSeq continues to improve, more samples can be multiplexed in a single run without sacrificing depth of coverage.

Finally, the advantages of PacBio SMRT sequencing are remarkable when the goal is sequencing, assembly, and forensic analysis of an unknown pathogen from a purified culture. As discussed above, PacBio SMRT technology allows for read lengths many times longer than that of other NGS technologies, at the expense of throughput. While the relatively high throughput of other NGS technologies provide high coverage of a genome, the short read lengths and amplification biases lead to fragmented assemblies wherever repeats or poorly amplified regions lie (Roberts, Carneiro, & Schatz, 2013). "Closing" these gaps to produce a high-quality draft genome is expensive and time-consuming, often requiring further sequencing and specialized library preparations. Long reads capable of spanning large repeats and poorly amplified regions drastically simplify the problem of finishing a genome assembly, and finished, "closed" bacterial genomes can now be readily produced using this data (C.-S. Chin et al., 2013). Draft (unfinished) genome assemblies are composed of many contigs, and may not necessarily represent the full biological potential of the organism in question. Furthermore, PacBio SMRT sequencing inherently measures the speed at which bases are added to the individual template molecules, and these kinetics allow for detection of methylated DNA bases, directly shedding light on functional information that can serve as a further forensic signature. While the low throughput of this technology prevents it from being ideal in all situations, some microbial forensics applications will benefit greatly from using this "3rd-generation" technology. As the read lengths, throughput, and portability of this technology improves, it may deserve a prominent spot in the microbial forensic analyst's toolbox, which will necessitate further bioinformatics R&D to make the most of this capability.

Metagenomics

NGS methods have been used extensively for whole genome analysis of single organisms, but only recently have methods become available for characterizing whole populations of microorganisms or for forensic use in decision-making. Dramatic improvements in NGS technology over the last several years (Wetterstrand, 2013) have enabled *metagenomics* – the application of sequencing to study DNA collected directly from environmental samples such as soil (Daniel, 2005), sea water (DeLong, 2005; Rusch et al., 2007), hospitals (Kembel et al., 2012), and human-associated habitats (Human Microbiome Project Consortium, 2012), without culturing or enrichment (Figure 3). Thus, a metagenome is the complete collection of genetic material recovered directly from an environmental sample, representative of the organisms present in that sample (bacterial, viral, human, animal, or otherwise). The National Research Council recently compared advancements in metagenomics to "a reinvention of the microscope in the expanse of research questions it opens to investigation" (National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications., 2007).

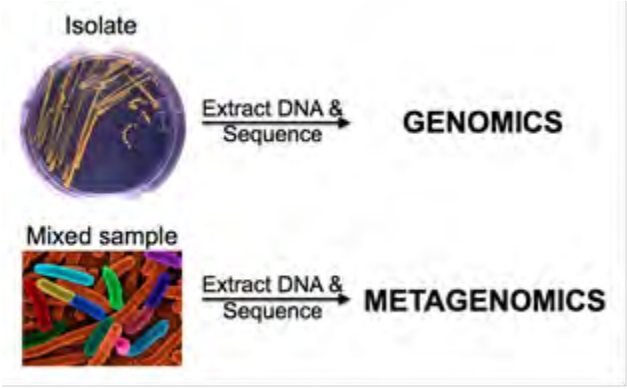

Figure 3: Genomics=sequencing a single genome. Metagenomics=sequencing DNA directly from a mixed environmental sample.

Analytical approaches for sequencing of a single organism from a pure culture have limited utility when a sample comprises a mixture of many organisms. Clone-based amplification can only confirm the presence of a known or suspected organism. BLAST searching individual reads is a computationally daunting task – the cost of a single BLAST search of metagenomic sequencing data can be approximately ten times the cost of sequencing (J. Wilkening, Wilke, Desai, & Meyer, 2009).

Most studies of microbial communities have used targeted amplicon sequencing of phylogenetically important marker genes, such as the 16S rRNA gene (Kuczynski et al., 2012). The 16S rRNA gene is commonly used because it is present in all living microorganisms, and it contains very slowly evolving regions that can be used to design broad-spectrum amplification primers. However, species-or sub-species-level resolution using 16S rRNA genes is problematic. The use of 16S rRNA gene profiling most often results in the identification of operational taxonomic units at higher level taxa, such as phyla (e.g. Firmicutes, Proteobacteria). These high-level classifications are rarely useful in microbial forensics. Furthermore, not all rRNA genes can be amplified with broad-spectrum primers, introducing bias and distorting the apparent microbial community structure. Finally, rRNA gene copy number can range from 1-15 in some bacteria (Z. M.-P. Lee, Bussema, & Schmidt, 2009), making rRNA gene-based approaches more qualitative than quantitative (Davenport et al., 2012). 16S rRNA approaches are often used as a proxy for studying the diversity of a microbial community, but sub-species and strains are often indistinguishable by their 16S rRNA gene, whereas they differ throughout the rest of the genome.

Next-generation metagenomics sequencing produces billions of short DNA sequence fragments, and an environment's microbial composition can be estimated from such data. Aside from identifying

potential biothreat agents in a sample, a metagenomic profile of all the organisms in a sample could be thought of as a particular sample's microbial "fingerprint" or "barcode" (Figure 4). Forensic attribution using shotgun metagenomics sequencing data on environmental samples requires methodology that can correctly assign sequencing output to individual microbial species, sub-species, or strain, at a level of resolution not achievable by probing the 16S rRNA gene alone. This can be achieved by analyzing NGS reads directly, analyzing metagenome assemblies, or a combination thereof.

***Unassembled read-based approaches for metagenomic analysis*:** One of the most widely publicized studies using a whole-genome shotgun metagenomics approach was the Human Microbiome Project (Human Microbiome Project Consortium, 2012). The HMP generated over 8 terabases of sequence data, which required development of an end-to-end quality control and data analysis pipeline (Gevers, Pop, Schloss, & Huttenhower, 2012). At the time, analyzing unassembled sequencing reads directly was considered too low resolution and not sufficiently accurate (Wommack, Bhavsar, & Ravel, 2008). However, with read lengths from Illumina's sequencing technology increasing from 76 bp to more informative >100 bp in the 1-2 year period between the pilot and production phases of the HMP (Gevers et al., 2012), direct analysis of unassembled sequencing reads became feasible.

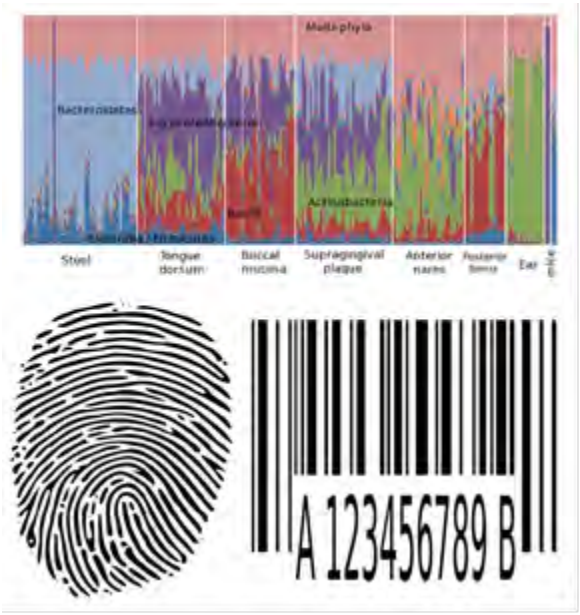

Figure 4: A sample's microbial community profile (top) can be thought of as its molecular fingerprint or barcode (bottom).

Methods for read-based taxonomic classification of metagenomics data fall into two broad categories: alignment-based and composition-based. Composition-based approaches rely on comparing signatures of short motifs from a query fragment to a reference genome – for instance, a particular GC content, gene and protein family content, or *k*-mer frequency and distribution (Thomas, Gilbert, & Meyer, 2012). Composition-based approaches include Phylopythia (McHardy, Martín, Tsirigos, Hugenholtz, & Rigoutsos, 2007), PhylopythiaS (Patil, Roune, & McHardy, 2012), Phymm (Brady & Salzberg, 2009), the Naive Bayes Classifier (Rosen, Reichenberger, & Rosenfeld, 2011), Sequedex (Berendzen et al., 2012), the Livermore Metagenomic Analysis Toolkit (LMAT) (Ames et al., 2013), and others.

Alignment-based approaches compare reads to a set of labeled reference genomes using a BLAST-based approach. One of the first such algorithms was the MEtaGenome ANalyzer (MEGAN) (Huson, Auch, Qi, & Schuster, 2007), developed in 2007, during the period in which short-read Illumina sequencing gaining acceptance. The MEGAN approach assigns reads to a taxonomy by first mapping reads to known genomes using BLAST, then running a simple lowest common ancestor algorithm to assign reads to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. Other much faster alignment-based methods have been developed more recently that use newer alignment algorithms (Davenport et al., 2012; V. K. Sharma, Kumar, Prakash, & Taylor, 2012), reduced databases (B. Liu, Gibbons, Ghodsi, Treangen, & Pop, 2011), or a combination of both approaches (Segata et al., 2012). Newer alignment methods developed for next-generation sequencing such as Bowtie (Langmead & Salzberg, 2012) can be hundreds to thousands of times faster than BLAST. A method used by the Human Microbiome Consortium called Metagenomic Phylogenetic Analysis (MetaPhlAn) uses a reference database reduction strategy (Segata et al., 2012). The reference database used by this method is compiled by starting with every prokaryotic reference genome available and removing sequences showing high similarity across distantly related taxa which would disrupt efforts to make unambiguously placed reads on a phylogenetic tree based on sequence similarity (Haft & Tovchigrechko, 2012). MetaPhlAn then uses either BLAST or Bowtie to search a reference database that is approximately 5% of the original size (meaning that the method has to search a much smaller space to match the query sequence to a reference genome). Other similarity-based read classification tools include MetaPhyler (B. Liu et al., 2011), CARMA (L. Krause et al., 2008), WebCARMA (W. Gerlach, Jünemann, Tille, Goesmann, & Stoye, 2009), IMG/M (Markowitz et al., 2012), and MG-RAST (F. Meyer et al., 2008).

Finally, other methods for direct taxonomic classification of sequencing reads use a combination of both composition and sequence similarity approaches, such as MetaCluster (Y. Wang, Leung, Yiu, & Chin, 2012), Rapid Identification of Taxonomic Assignments (MacDonald, Parks, & Beiko, 2012), and PhymmBL (Brady & Salzberg, 2011), which integrates interpolated Markov model-based composition-directed taxonomic predictions with BLAST-based homology results. These hybrid methods have been shown to represent improvement over the performance of their individual components, but not over other optimized purely homology-based approaches (Haft & Tovchigrechko, 2012; Segata et al., 2012). A more comprehensive review of sequence classification methodology and software is presented elsewhere (Bazinet & Cummings, 2012).

***Metagenome assembly*:** From a microbial forensics perspective, a read-based approach often provides a quick, low-resolution answer to the question "who's here?". It may not, however, address the follow-up questions, such as "who are they most related to?" and "what are they doing?" that may be of importance in a microbial forensics investigation. If the goals are either to recover the full genome of an uncultured organism in an environmental sample for fine-grained taxonomic classification, or to reconstruct coding sequences for functional characterization of the genes present in the environmental sample, then assembly of short reads into longer contigs and scaffolds is necessary (Thomas et al., 2012).

A number of methods have been described for *de novo* assembly of NGS reads without a reference genome (see above). Most of them rely on de Bruijn graph representations in which nodes are "words" of k nucleotides (k-mers) that are observed in the sequencing data, and edges between nodes are drawn when the k-mers in those two nodes appear consecutively in a read (@Compeau2011c; W. Zhang et al., 2011). Genome assembly proceeds by traversing the graph – that is, finding the Eulerian cycle that visits all the nodes in the graph exactly once (Flicek & Birney, 2009). The assembly process is

computationally challenging with short sequencing reads, and numerous methods have been developed to assess the quality of the final product (Narzisi & Mishra, 2011; Vezzi, Narzisi, & Mishra, 2012) as illustrated by genome assembly competitions to crowdsource evaluation of the differentmethods (K. R. Bradnam et al., 2013; Earl et al., 2011; Magoc et al., 2013).

Many assemblers exist for d*e* nov*o* reconstruction of single genomes, but working with reads from abundance-skewed populations of molecules, such as a metagenome, poses greater and more complex challenges than with single-genome assembly. This is because the relative abundances of species in a microbiome are not uniform (J. R. Miller, Koren, & Sutton, 2010). Several tools have recently been developed for d*e* nov*o* transcriptome assembly (expression is also highly skewed between genes) (Grabherr et al., 2011; Schulz, Zerbino, Vingron, & Birney, 2012). Metagenome assembly, on the other hand, poses additional challenges, such as substantial variation in abundance at the species level, and genetic variation between closely related species. Using a single genome assembler for metagenome data results in highly abundant species being identified as "repeats" in a single genome. Further, genetic variation between species results in "branches" in the de Bruijn graph, making it more complex and difficult to traverse. Very recently, several de Bruijn graph-based methods have been developed to account specifically for the heterogeneous mixture of organisms present in a metagenomic sample. Meta-IDBA partitions the de Bruijn graph into components of different species, then captures the variants in genomes of subspecies from the same species by multiple alignments and represents the genome of a species using a consensus sequence (Y. Peng, Leung, Yiu, & Chin, 2011). MetaVelvet is an extension of the Velvet assembler for d*e* nov*o* assembly of mixed-sequence reads from multiple species in a microbial community (Namiki, Hachiya, Tanaka, & Sakakibara, 2012). Similar to Meta-IDBA, MetaVelvet decomposes a de Bruijn graph from mixed short reads into individual sub-graphs and builds scaffolds based on each decomposed subgraph as an isolate species genome. Recent work has shown that adding color to de Bruijn graphs enables recognition of genomic variants from simultaneously assembling multiple genomes d*e* nov*o* (Zamin Iqbal, Caccamo, Turner, Flicek, & McVean, 2012; Z. Iqbal, Turner, & McVean, 2012). Ray Meta is a metagenome assembler based on uniquely colored *k*-mers capable of assembling billions of metagenomic reads from 1,000 bacterial genomes of uneven proportions under reasonable CPU and memory constraints by employing massively parallel distributed computing that isn't supported in other metagenomics assembly algorithms (S. Boisvert, Raymond, Godzaridis, Laviolette, & Corbeil, 2012).

Extremely deep sequencing is necessary for capturing the full extent of microbial diversity within highly complex environments; this often results in very high memory requirements for assembly algorithms that exceed the capacity on most modern computers (Fierer et al., 2012; Hess et al., 2011; Pell et al., 2012; J. Qin et al., 2010). To address the scalability problems facing metagenome assembly, new methods have been developed. One of these uses a probabilistic representation for storing de Bruijn graphs in memory based on Bloom filters (Pell et al., 2012). Using a fixed memory probabilistic data structure and de Bruijn graph partitioning, this method is able to store and traverse DNA de Bruijn graphs using 20-to 40-fold less memory than other common de Bruijn graph-based assemblers. Another method – digital normalization – relies on data reduction prior to assembly (Brown, Howe, Zhang, Pyrkosz, & Brom, 2012). Because extremely high coverage is necessary to capture low-abundance organisms in a mixed sample, there is an over-sampling of reads from high=abundance organisms present in a mixed sample. This sequence duplication leads to increased computing time and memory requirements. Digital normalization reduces computational requirements prior to assembly by removing the majority of (redundant) reads without affecting the *k*-mer content of the dataset, and by simultaneously eliminating sequencing errors, which would otherwise add to the memory requirement for assembly. A drawback to employing a data reduction strategy, such as digital normalization, is that

all coverage information is abandoned. If a forensic application only requires knowing what is present in a sample and abundance information is either unimportant or can be recovered in some other way after a first-pass analysis, digital normalization provides a much more rapid assembly with fewer computational requirements.

Task 2: Rapid discrimination of natural, accidental and deliberate infectious disease outbreaks

From a policy and decision-making standpoint, it is critical to determine whether an outbreak is due to *natural* circumstances, *accidental* release of a cultured or engineered organism, or *deliberate* introduction of a known pathogenic organism. Under these circumstances, the ultimate goals are identification of the agent responsible for an outbreak, and attribution of both its biological and geopolitical origins to a single source at the exclusion of all other sources with a high degree of certainty. The resultant information is central to making correct decisions that are required in response to the outbreak, such as whether a public health emergency should be declared. Genome sequencing offers the potential for rapidly providing ultra-high resolution signature data, thus enhancing the likelihood of meeting criteria described in the International Health Regulations (World Health Organization, 2008). U.S., E.U. law and World Trade Organization regulations require quick responses with a scientific basis to avoid needless disruption of trade and travel relating to a potential Public Health Emergency of International Concern. Genome sequence information contains valuable clues on what scientific and legal mechanisms should be used to limit transmission and contain the outbreak and even whether to trigger a response under the Biological and Toxin Weapons Convention.

Below we highlight current epidemiological and molecular considerations for making a determination of whether an infectious disease outbreak is the result of a deliberate, accidental, or natural release. We then present several case studies highlighting the advances that NGS has contributed to the understanding of the molecular epidemiology of outbreaks from a public health prospective. We conclude with a series of challenges, opportunities, and recommendations for future investment that will enable the microbial forensics field to realize the full potential of NGS in contributing to a determination of deliberate, accidental, or natural disease outbreaks.

Epidemiological factors to discern between deliberate, accidental, and natural outbreaks

A number of authors have described epidemiological indications that an outbreak may be caused by bioterrorism (Dembek, Kortepeter, & Pavlin, 2007; Grunow & Finke, 2002; Radosavljevic & Belojevic, 2012; Rotz & Hughes, 2004; Treadwell et al., 2003). Dembek et al. list eleven “clues” supporting the likelihood of an intentional attack (Dembek et al., 2007):

1. Unusual event: A highly unusual event with large numbers of casualties occurs with no plausible natural explanation.
2. Increased morbidity/mortality: The outbreak results in higher morbidity or mortality than is expected.

3. Uncommon disease: Most infectious diseases have predictable geographical, environmental, and host endemicity. Outbreaks in uncommon locales may represent an unnatural outbreak.
4. Point-source outbreak: Intentional outbreaks would likely follow a point-source outbreak curve (Inglesby, 1999), with a quick rise in cases, followed by a plateau and an acute drop.
5. Multiple concurrent epidemics: Intentional releases may occur at various locations at once.
6. Lower attack rates in protected individuals: If the attack rate is lower in individuals protected from biological incident or attack by vaccination, antibiotics, wearing personal protective equipment or living in an air/water barrier isolation unit, this may be indicative of an unnatural outbreak.
7. Dead animals: Dead animals may indicate intentional release of zoonotic pathogens that also affect humans (Meselson et al., 1994).
8. Reverse spread: Spread of disease from people to animals may indicate an unnatural disease outbreak.
9. Unusual disease manifestation: Atypical clinical manifestation may indicate an unnatural spread of disease (e.g., inhalation anthrax when most natural anthrax cases are cutaneous).
10. Downwind pattern: A disease pattern consistent with the discharge of the pathogen from a single point followed by its subsequent distribution by air or water currents may indicate an outbreak of unnatural origin.
11. Direct evidence: The "smoking gun," such as a letter filled with anthrax spores (Dalton, 2001) clearly indicates deliberate release of a pathogen.

Appearance of any of these clues during an outbreak should prompt consideration that an accidental or deliberate release of a bioweapon may be involved. However, it is important to note that no single one of these clues is necessary and/or sufficient to prove that an outbreak was intentional, and that many of these clues have been observed in at least one investigation of a naturally occurring outbreak.

Others have developed epidemiological approaches that yield a quantitative assessment of whether an outbreak is intentional, accidental or natural (Grunow & Finke, 2002; Radosavljevic & Belojevic, 2012). Since these tools are based on incomplete evidence and qualitative data, they are necessarily imprecise. Neither test has been fully validated or applied in the course of an outbreak. However, both could provide guidance to public health officials in the early stages (or in a later retrospective assessment) of an event for making decisions such as whether and when to include law enforcement in the investigation. Dembek et al. applied the Grunow & Finke test retrospectively to a number of case scenarios and found it accurately identified the situations most likely to have resulted from bioterrorism, but failed to pick up deliberate attacks that involved common pathogens (Dembek et al., 2007). Moreover, the difficulty with this tool is that considerable investigation may be required before a valid assessment can be made, and by then, crucial leads may be lost or new events initiated.

In that Grunow & Finke's approach is felt to be too slow for use in attacks that involve mass casualties, Radosavljevic & Belojevic created an alternative method that uses similar epidemiological clues but weights factors differently, enabling a more rapid assessment of the nature and etiology of an outbreak. Application of the Radosavljevic & Belojevic test to the North American outbreak of swine flu in 2009, the Kosovo tularemia outbreak of 2000, and the Sverdlovsk anthrax release of 1978 yielded conclusions, which were consistent with those obtained after final analysis of these events (Radosavljevic & Belojevic, 2012). Applying a tool retrospectively is different, however, than using it in real time. The difficulty with both the Grunow & Finke and Radosavljevic & Belojevic tools is that it is difficult to anticipate the availability and accuracy of data.

Using NGS to characterize an outbreak: case studies

The previous section dealt with epidemiological clues that inform whether an outbreak may be unnatural. In reality, public health officials often need to act long before all of this epidemiological information can be assembled. Thus, although epidemiological tools and other conventional indicators will continue to play a crucial role in the characterization of a suspicious outbreak, concurrent application of genome sequencing and other molecular characterization will generally yield more rapid and accurate results. Radosavljevic & Belojevic found that of all the variables they examined, identification of the etiologic pathogen was the most important component of the scoring system and concluded that pathogen genome sequencing should be applied as soon as any suspicious indicator is identified (Radosavljevic & Belojevic, 2012). Here we review several recent examples of where genome sequencing was used during an infectious disease outbreak. While the following examples are taken from the public health literature, the salient features related to microbial forensics are noted.

Case study: Shiga toxin-producing *E. coli* O104:H4

E. coli are ordinarily commensal, enteric organisms but several, such as enteroaggregative *E. coli* (EAEC) strains and Shiga toxin-producing enterohemorrhagic *E. coli* (EHEC) strains possess virulence factors that enable them to be pathogenic. Shiga toxin can cause both local damage in the gut (resulting in bloody diarrhea) and, after entry of the organism or its products through the gut wall, can damage the kidney resulting in hemolytic uremic syndrome (HUS) (Kaper, Nataro, & Mobley, 2004).

In May-June 2011, over 4,000 cases of bloody diarrhea and 850 cases of HUS, including 50 deaths, were reported initially in Germany and subsequently in France. These illnesses and fatalities were subsequently determined to be caused by a strain of Shiga toxin-producing E. col*i* serotype O104:H4. Historically, the O104:H4 strain has not been associated with a high rate of HUS (Bae et al., 2006; Mellmann et al., 2008), but the proportion of infected patients who went on to develop HUS and other complications in this outbreak was unexpectedly high (Frank et al., 2011; Jansen & Kielstein, 2011). Importantly, the German and French outbreak isolates were indistinguishable using *conventional* molecular methods, including serotyping, multilocus sequence typing (MLST), virulence gene content typing, rep-PCR, pulsed-field gel electrophoresis, optical mapping, and antimicrobial susceptibility testing (G. Gault et al., 2011; P. Mariani-Kurkdjian, Bingen, Gault, Jourdan-Da Silva, & Weill, 2011). Clearly, a higher-resolution molecular technique was needed to provide a more detailed understanding of the molecular evolution of this emergent pathogen.

In May of 2011, an "open-source genomics" approach was developed to characterize the genomic signature of this novel pathogen using high-throughput benchtop NGS, open-data release, and rapid outsourcing ("crowdsourcing") of the analysis to interested bioinformaticians worldwide (H. Rohde et al., 2011). An isolate from one member of an infected family was sequenced using a benchtop NGS platform (Ion Torrent – this pre-dates the release of the Illumina MiSeq). As soon as the sample was sequenced, the data was released into the public domain (https://github.com/ehec-outbreak-crowdsourced/), eliciting a wave of curiosity-driven analyses from interested bioinformaticians on four continents. Within one day of data release the genome had been assembled; within 2 days it had been assigned to a sequence type; within five days strain specific diagnostic primer sequences were released, and within a week, 24 reports were submitted to the open-source analysis wiki. Soon afterward, the pathogen was sequenced to a high depth of coverage on an Illumina HiSeq platform (H. Rohde et al., 2011) and with long reads using PacBio SMRT chemistry (Rasko et al., 2011). Further analyses revealed that this particular outbreak strain contained a prophage encoding Shiga toxin, a rare

combination of virulence factors, a plasmid that encodes several enteroaggregative *E. coli*-specific virulence factors, and another plasmid encoding an extended spectrum beta-lactamase (antibiotic resistance) (Grad et al., 2012; Rasko et al., 2011; H. Rohde et al., 2011). This rare cocktail of virulence factors, antibiotic resistance, and several of Dembek's epidemiological indicators (Dembek et al., 2007) listed above may have aroused suspicion that this outbreak resulted from an unnatural release. However, phylogenetic trees created from the whole genome sequences obtained from the O104:H4 outbreak and related strains supported the hypothesis that the Shiga toxin-producing outbreak strains evolved from an enteroaggregative *E. coli* O104:H4 that recently acquired the toxin-encoding phage via natural lateral gene transfer. Therefore, the most plausible interpretation of the accumulated data regarding this episode is that the source was a Shiga toxin-producing O104:H4, which arose via a naturally occurring exchange of genetic material, not deliberate or accidental release of an engineered bioweapon. Deliberate introduction of Shiga toxin-encoding prophage would likely have shown other features, such as the presence of cloning vectors (containing known restriction sites) or other genomic features that pointed to a laboratory *E. coli* strain.

Although the outbreak was first recognized in relation to salad products eaten in Northern Germany (and later in a separate but related outbreak in France), it quickly involved most of Europe. Within two weeks, twelve European nations had reported outbreaks that constituted a public health emergency of international concern and resulted in severe economic consequences that crossed international borders. Before a sophisticated analysis identified the source, many countries acted proactively to remove cucumbers and tomatoes from store shelves, Russia issued a trade ban on Spanish cucumbers, and the Spanish cucumber and vegetable market suffered significant losses (estimated at 200m euros a week) (Tremlett, G, Pidd, 2011). This case demonstrates the vulnerability of our food supply and why unusual outbreaks involving endemic microbes must be taken as seriously as those involving the typical agents of bioterrorism. Although there did not appear to be criminal intent involved, the resulting, complicated economic and legal issues documented the importance of having and using sophisticated microbial forensic techniques to aid in such an investigation and any subsequent legal proceedings.

Finally, as noted above, NGS methods have been used extensively for single-organism whole-genome analysis, but metagenomics sequencing of all DNA in a sample without purification and amplification is now practical. In a retrospective study, metagenomic stool samples collected from patients during the 2011 O104:H4 outbreak in Germany were sequenced. The resulting data illustrates the capability and soon-to-be-overcome limitations of metagenomic approaches, highlighting areas for future attention, such as sample collection and sequencing depth that need to be further developed in order for this approach to play a key role in the genetic characterization of a pathogen during an outbreak (Loman et al., 2013). Here, investigators used total DNA directly isolated from clinical samples to prepare sequencing libraries which were then run on both the benchtop Illumina MiSeq and multiplexed HiSeq instruments. Encouragingly, the researchers were able to assemble out of this mass of data a full draft genome of the causative pathogen. Another arguable success is that they were able to detect sequences from the pathogen in 67% of the samples – a sensitivity that is low for a clinical diagnostic, and may yet be too low for forensic attribution purposes. Notwithstanding these results, the study found many potential pathogens in any particular sample. To identify the causative agent, they had to determine which species were present using only microbial sequences that were present in at least 20 patients, but not in any healthy controls. The investigators concluded from their studies that metagenomic data alone currently cannot be used to infer a causal link between detecting a pathogen *in a single sample* and the coincident occurrence of disease, and that it is essential to collect multiple "suspect" and "background" samples to determine if there is a statistically significant link to support a

causal relationship (Loman et al., 2013).

Although not a deliberate act, this example illustrates the vulnerability of food products to malevolent tampering, and the widespread international economic consequences that can occur even from limited product contamination. Outbreak strains were indistinguishable from other strains using conventional molecular techniques, and only whole-genome sequencing could provide the necessary high-resolution analysis to identify the origins of this novel pathogen. A very recent retrospective metagenomic study illustrates the current capabilities and limitations of using metagenomics to infer the cause of an outbreak, and is an example of a sophisticated bioinformatics analysis that could provide such an interpretation. With better, faster bioinformatics analysis and collection of reference samples, it will be possible to attribute a source in a shorter time and with a higher degree of confidence (see *The path forward* section below).

Case study: Methicillin-resistant *Staphylococcus aureus*

While not thought of as a biological weapon, Methicillin-resistant *Staphylococcus aureus* (MRSA) is a major public health concern, and is endemic in both civilian and battlefield hospitals worldwide. In 2005, more than 90,000 cases of invasive MRSA infections occurred in the United States, resulting in greater than 18,000 deaths (Klevens et al., 2007), and MRSA infection typically doubles the cost and length of stay of hospital inpatient visits. Furthermore, community-acquired MRSA is known to be common among military recruits (Zinderman et al., 2004), and some strains of MRSA have been recently noted as causes of bloodstream infections at military medical centers (Sherwood, Park, Robben, Whitman, & Ellis, 2013), making MRSA-transmission network reconstruction a priority for initiatives in force protection and warfighter healthcare improvements.

Typically molecular genotyping methods are used to complement epidemiological investigation by infection control and surveillance teams to reconstruct transmission networks, such that current outbreaks can be attributed to a source and future risks mitigated. However, because MRSA has a clonal population structure with only a few bacterial genotypes that cause the majority of outbreaks (Enright et al., 2002), most conventional molecular techniques like multilocus sequence typing (MLST) or pulsed-field gel electrophoresis (PFGE) do not have the necessary discriminatory resolution to differentiate between outbreak and non-outbreak (endemic) strains. The same clonal nature of organisms and limited resolution of conventional molecular methods are true for many potential agents of biological warfare or terrorism, including *Bacillus anthracis*, *Yersinia pestis*, *Francisella tularensis*, *Coxiella burnetii*, and *Burkholderia mallei*. These obstacles can, however, be eliminated with the introduction of NGS and the bioinformatics methodology to support it. Here we review a recent high-profile example of a nosocomial MRSA outbreak in which NGWGS was used to reconstruct transmission dynamics and determine that the cause of the outbreak was accidental. While these high-profile cases are taken from the academic literature on an organism of public health concern, the same principles and procedures can be used to investigate an outbreak involving a potential bioweapon in order to attribute a source and determine whether the outbreak is natural, accidental, or deliberate.

NGS can provide a higher level of resolution than conventional molecular assays as has been shown in the *E. coli* study above. Benchtop NGS was used for the first time in 2012 to distinguish between MRSA isolates that were associated with an outbreak in a neonatal intensive care unit and endemic MRSA isolates (Köser et al., 2012). However, this example used NGS retrospectively. In one case reported earlier this year from the Rosie Hospital (part of the Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK), NGWGS was used *in real time* to complement the

epidemiological investigation of several MRSA clusters in the special baby care unit (SBCU) (Harris et al., 2013).

Surveillance and investigation by the infection control team identified three clusters of MRSA sharing an identical antibiogram (antibiotic susceptibility profile), where the outbreaks were separated by gaps of 17 and 33 days during which no MRSA cases were reported. The gaps meant that the infection control team was not able to conclude whether a putative outbreak extended over the entire period, or if these reflected a second or third pathogen. Sequencing MRSA cultures from these samples on a benchtop NGS instrument (Illumina MiSeq) revealed that the strains were clonally related and represented a new sequence type. By incorporating NGS into a prospective epidemiological investigation, the infection control team uncovered new transmission pathways from babies to mothers, from mothers-to-mothers in the postnatal ward, and from mothers and partners to other healthcare workers. Importantly, the re-emergence of this particular outbreak strain (as determined by NGS) after a "deep clean" was a clue that the re-emergence was likely not "natural", and led the team to hypothesize that a member of the hospital staff may be carrying and transmitting the same MRSA strain. This concept was supported by the phylogenetic tree structure of the SBCU isolates, which did not show a signature of sequential transmission, but instead hinted at repeated introduction from an external source (Harris et al., 2013). After screening 154 SBCU staff members, one individual was positive for MRSA, which was confirmed as the novel strain type by NGS. The carrier was later relieved from clinical duties while undergoing MRSA decolonization therapy. By incorporating NGWGS into this investigation, the infection control team was able to (1) determine the outbreak was caused by a new sequence type, (2) attribute separate events of a re-emergent pathogen to a likely external source, and (3) determine that the cause was accidental transmission by a healthcare worker. The authors also noted that the cost of this outbreak was over 1000 times the cost of sequencing a single MRSA isolate. It is also important to note that the attribution of multiple outbreak strains to a single external source was made using an expert's bioinformatic analysis of the sequence and phylogenetic data (refer to the section below: "*Don't eliminate the bioinformatician...").

Other case studies

WGS has been used recently for a number human pathogens, including the ones referenced and several others (Bryant et al., 2013; C.-S. Chin et al., 2011; Gardy et al., 2011; Grad et al., 2012; Harris et al., 2013; 2010; Köser et al., 2012; Rasko et al., 2011; Roetzer et al., 2013; H. Rohde et al., 2011; T. M. Walker et al., 2013). A severe epidemic of cholera in Haiti in 2010 infected more than 90,000 people, killing over 2,000. Genome sequencing was performed on several Haitian *Vibrio cholerae* isolates, as well as strains that caused previous outbreaks in South America and South Asia, showing that the 2010 Haitian strain was nearly identical to the South Asian "seventh pandemic" strains, indicating a likely introduction of this particular cholera strain to Haiti from a distant geopolitical source (C.-S. Chin et al., 2011; Reimer et al., 2011). Next-generation whole-genome sequencing was recently used to track a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae*, which also has a clonal population like MRSA, making it difficult to discriminate between strains to make an attribution using conventional molecular techniques. Recently in British Columbia next-generation whole-genome sequencing in combination with social network analysis was used to investigate a three year tuberculosis outbreak. The sequencing data overlaid with the social-network analysis showed that a socioenvironmental factor, most likely crack-cocaine use, triggered the simultaneous expansion of two existing genetic lineages of *M. tuberculosis*, where traditional VNTR genotyping and epidemiological interviewing failed to identify a source. We direct the reader to the primary sources referenced herein as well as those reviewed elsewhere (Le & Diep, 2013).

Legal issues to consider

Surveillance

Indicators of the release of an infectious agent include the occurrence of disease in humans as well as data from surveillance systems. There is a need to coordinate public health, law enforcement, and national security activity under a common umbrella and during this concerted action, to employ the recent advances in WGS as applied to microbial forensics. It is essential that the details obtained during bio-surveillance be integrated into the analysis of samples in order to obtain the maximum amount of useful information.

The U.S.'s current strategy for bio-surveillance is an "all-hazards" approach, so that potentially dangerous incidents are monitored through the same surveillance systems regardless of whether they are believed to result from deliberate or accidental release or natural occurrence. Since 2001, environmental and informational surveillance systems have been greatly expanded in type and geography. Outbreaks can be detected using biological agent detectors. An example of these is the Bioagent Autonomous Networked Detector (BAND) Project, which involves air sampling in at-risk urban areas at least once every three hours (Ervin, Hultgren, Rhyne, & Ward, 2010). The BAND Project focuses on pathogens such as Bioterrorism Category A agents: anthrax, botulism, plague, smallpox, tularemia or a viral hemorrhagic fever. These surveillance systems involve several agencies that have cooperated to provide sample libraries for detecting potential threats and ruling out background material. They employ protocols consistent with microbial forensic applications, have sophisticated sampling algorithms, and involve multiple national security and law enforcement agencies. Programs to increase water sampling have been initiated, but are technologically much more difficult. Despite increasing cooperation between state, federal, and international agencies, the limits of physical surveillance methods make it likely that an attack will first be detected through epidemiological indicators, rather than environmental sensors (Lipsitch, Finelli, Heffernan, Leung, & Redd, 2011).

Electronic health records throughout the U.S. can be used as a form of surveillance to identify potential epidemics in their early stages. This surveillance includes new reporting requirements and coordination both nationally within state and federal frameworks and internationally. While the Privacy Rule permits disclosure of private health information without authorization for public health surveillance, investigations, and interventions, the disclosure of this information has legal and logistical consequences, since criminal prosecution and national security interests are not always the same as those of public health (V. Sutton, 2005). Most developed states and countries have privacy statutes that dictate how patient records may be used. These laws, which differ considerably among states and countries, need to be taken into consideration prior to sample collection. Since 2001, a number of protocols have been developed/proposed to deal with these conflicts but this has not been addressed in many of the states. Even in states that have dealt with these issues it is important to recognize the impact of new technologies (V. Sutton, 2005).

Privacy and patient records

In the United States, HIPAA (the Health Insurance Portability and Accountability Act) allows free access to patient records held by "covered entities" (which include various health plans, health-care clearinghouses and providers that transmit health information electronically) for legitimate public health purposes, but more limited access for judicial proceedings. For the latter, a court order or other legal order (e.g. a subpoena) is necessary for the release of the records and notice must be provided to

the person whose records are sought. If environmental samples might include evidence to incriminate the individual whose health records are sought, additional state and federal criminal evidence safeguards may exist.

There is a potential loophole in that HIPAA does not block public health officials from sharing patient records with law enforcement officials (because public health entities are not "covered entities" under HIPAA). There could, however, be serious political and public trust consequences if the public health officials were perceived as being used to "spy" on individuals for law enforcement. The full scope of such an action will depend on the national security protocol that is invoked, and while, in that setting, a covered entity may disclose patient records, it is not required to do so.

As authorized by the National Security Act and the USA Patriot Act, or Executive Order, covered entities can release patient records to authorized federal officials for the conduct of lawful intelligence, counter-intelligence and other national security proceedings. A full debate on the limits of emergency powers and how patient privacy may be affected has not occurred, and there are significant gaps in our understanding of how patient records may be used in a public health emergency. The initial discussions on HIPAA exceptions have focused on the patient records of a potential perpetrator, and while discussions have started on the creation of large bioterrorism databases of patient records, these debates have not yet resulted in any clear directives.

State and federal laws may apply to patient record access. If a public health emergency is declared, states that have passed legislation incorporating the Model State Emergency Health Powers Act have sweeping powers to deal with it. This would allow significant access to patient records by authorized personnel, but as different states may have different limitations and definitions of a public health emergency, access would need to be dealt with on a case-by-case basis. The full extent of coordination among public health officials, law enforcement and national security officials will be context-dependent and has not yet been fully tested. In addition to the challenges of patient-record access, their utility in routine surveillance will be affected by heterogeneous institutional facilities, resources and reimbursement agreements, as well as differences in clinical decision-making which may result in inconsistent sampling.

The "all-hazards" approach may be an adaptable strategy to safeguard public health of the nation. In natural pandemics and the deliberate or accidental release of infectious agents, the initial casualties may be seen first in hospitals or other health care facilities. Information indicative of the early phase of an outbreak will be available initially to public health officials who are trained to gather the epidemiological evidence to contain disease and treat affected individuals. However, these public health officials are often not knowledgeable about nor focused on determining the source of the agent, identifying a possible perpetrator, or collecting and preserving evidence for legal proceedings. It is therefore essential that the leaders in this area have knowledge and experience to recognize the potential for an accidental or deliberate release and the authority to initiate an alert and associated response. Access to the current technologies used for microbial forensics, including next-generation whole genome sequencing and bioinformatic interpretation, would be beneficial in the investigation of a routine outbreak as well as one that may involve criminal activity and/or national security.

The path forward

The case studies presented above illustrate a recurring theme – namely, that NGWGS provides a fast, inexpensive and high-resolution method for discriminating among very similar strains of an extant or emerging pathogen, in a setting where conventional methods fail. This information can assist

in making a determination of whether an infectious disease outbreak is the result of a deliberate, accidental (as in the MRSA case), or natural (as in the *E. coli* case) event. The first step in distinguishing among deliberate, accidental, and natural is deciding whether an outbreak is *unusual* and then reconstructing potential transmission pathways. The cited cases illustrate how this concept is currently being used in public health. Below we draw out a plan for a path forward to future deployable capabilities for using NGS in microbial forensics applications. We return to the deliberate, accidental, or natural distinction in the *Conclusions* section.

Invest in bioinformatics

In a perfect world for forensic attribution, an enemy combatant or other violent actor will leave behind petri dishes inoculated with a pure culture of the agent being used as a bioweapon. DNA extraction and sequencing, and bioinformatic analysis would be relatively straightforward (and may not even be necessary, given other cheaper and faster technologies – see below). However, this is unrealistic, and forensic investigators must be prepared to deal with mixed, contaminated, or otherwise complex samples – such as from air filters, surface swabs, biological/anatomical sites, heavily contaminated with host cells, or at worst, very highly complex samples such as from soil. As illustrated above with the retrospective *E. coli* metagenomics study, applying genomic technology to a mixed sample of unknown origin and composition has great strengths when paired with proper bioinformatic/epidemiological analyses. At the same time, this scenario illustrates the significant limitations that arise when expert bioinformatic analysis is not available.

Relevant concentrations of bioweapon organisms may be too low for separation of signal from background noise or contamination. Solving this problem will require: (1) a depth of sequencing coverage that is in excess of what current technology is capable of producing under realistic time and budget constraints; (2) rapid and memory-efficient bioinformatics tools for attributing sequences to source organisms at the species, sub-species, and strain levels; (3) complete and frequently updated reference databases for comparison, analysis and detailed discussion; and (4) multiple replicates of the query "suspect" sample, as well as adequate background/endemic sampling so as to enable a statistical distinction between a threat and "normal" background. The first of these (#1) will naturally become available as sequencing technology proceeds at its current faster-than-Moore's-Law pace. Development of bioinformatics tools (#2) to process and analyze the next generation of sequencing data, will also naturally progress in the open-source community as the technology enables new applications. However, goals for methodology development in the open-source, academic community are not particularly geared toward forensic attribution. General objectives, such as improved clinical diagnostics, are similar, but more resources should be directed specifically toward development of bioinformatics methodology as it is concerned with microbial forensics. In addition, these approaches must be translated into robust, field-deployable technologies that can be used and interpreted by forensic scientists with minimal support from professional bioinformaticians. Reference databases (#3) are discussed in more depth below. Adequate replicate "case" and "normal" background sampling (#4) is a requirement that to some degree has always been recognized (robust statistical analysis cannot be performed with n=1), but its application to each situation will likely vary. Luckily, as sequencing technology becomes increasingly rapid and inexpensive, understanding endemicity on demand is not beyond the realm of possibility, and should be considered in any future forward-deployed capability.

The determination of whether an outbreak is deliberate, accidental, or natural hinges in part upon the molecular evidence captured from the sample – in the case of using NGS, from the bioinformatic analysis of the sequence data. A path forward to deployable capability must include a robust

framework for benchmarking, validation, and workflow management. Many reliable tools already exist for NGS analysis, performing functions such as quality control, assembly, classification and binning, computational subtraction of host sequences, and abundance estimation (discussed above in the *Bioinformatics* section). Before determining whether an outbreak is likely deliberate, accidental, or natural, a forensic analysis must be able to identify *what* is in the sample and specify a certain *level of confidence* based on available evidence. Benchmarking and validation are essential in the development of methods for microbial forensics to enable generation of reliable and defensible results that can be used by lawmakers and policymakers when responding to perpetually increasing force protection and national security threats (Budowle et al., 2008). In that the validation guidelines in Budowle *et al* were written more than five years ago, a group of subject-matter experts should be commissioned to update these recommendations, in light of contemporary molecular tools like NGS. Finally, workflow management and reproducibility are critical to the successful development and validation of a forensic analysis solution involving NGS. The sequencing itself, as well as all the bioinformatics processing that comes downstream, involves many moving parts, and reproducible, modular, robust tools for workflow management are indispensable. Some already exist for workflow management, from general workflow management tools like Taverna (Oinn et al., 2004), and bioinformatics-specific workflow management frameworks like Galaxy (Goecks, Nekrutenko, & Taylor, 2010), to metagenomics-specific workflow management and data analysis pipelines like MetAMOS (T. J. Treangen et al., 2013). There are, however, additional issues on "bioinformatics pipelines", which are discussed below.

Don't eliminate the bioinformatician – invest in bioinformatics training & support

What follows is deliberately provocative. The notion of a "bioinformatics pipeline" is inherently flawed. Consider the following: running an "automated cell biology pipeline" to understand the molecular pathogenesis of an infectious disease, or running an "automated intelligence analysis pipeline" to interpret signatures and assess a threat. Both are examples of complex scientific and analytical procedures that are overly simplified *ad absurdum*. Similarly, it is unreasonable to expect to completely automate *in silico* the science of managing and analyzing enormous amounts of data and interpreting molecular evidence in conjunction with sampling considerations, epidemiological evidence, and microbiological observations. The concept of a "bioinformatics pipeline" falsely implies that it is possible to automate navigation through many potential paths of a complex workflow, hard-coding many difficult decisions and interpretations along the way.

Fortunately, molecular biologists can be successfully trained in bioinformatics, quantitative methods, computational capability, and data literacy, to the degree that would enable the scientist to manage large amounts of data, perform analyses with existing tools, and interpret data critically in order to reach a statistically sound and robust conclusion. One of the authors of this report has published and actively maintains a collection of in-person and online training resources for bioinformatics, genetics, molecular evolution, statistics, and computational skills, such as Linux/Unix, scripting/programming, version control, databases, etc. (S. Turner, 2013). All of those resources as well as many others not covered therein are aimed at increasing the computational and quantitative skill sets for scientists who do not have formal training in these areas. Many, if not most, bioinformaticians were originally molecular biologists, who only later acquired the interest and skill to become competent in bioinformatics. Investment should be made toward informatics training for microbial forensic scientists and capability operators, such that the front-line forensic specialists are empowered to make robust interpretations quickly.

Reduce sample-to-answer time: using NGS to design faster & cheaper assays

Another opportunity/application for NGS in microbial forensics is using it to identify microbial genomic DNA sequences that correlate strongly with clinically important phenotypes, in order to develop a PCR assay that could quickly determine the host range, virulence, and antibiotic susceptibility of an organism. Complete genome sequencing of an organism and in-depth characterization of a metagenomic profile both enable analysis of the microbial genetic signature of a sample the highest possible resolution. However, NGS is still expensive relative to some conventional assays, can be more time consuming, and is less sensitive relative to conventional PCR assays, making it (arguably) less than optimal as a "front line" diagnostic tool to use for targeting and detecting the presence of known threats.

An area that should be prioritized for further research and development is the sequencing and bioinformatics methodology to determine very quickly DNA sequences that accurately predict clinically important phenotypes, like virulence and antibiotic susceptibility. Field application would then only require rapid and sensitive PCR amplification of the informative genetic segments. A recently described method is capable of doing this using an i*n* silic*o* optical mapping-like approach (B. G. Hall, Cardenas, & Barlow, 2013). This approach was shown experimentally to be able to distinguish very highly related Shigella, EHEC, and non-EHEC E. col*i* at 99% accuracy. However, this method currently only works for completed, "closed" genomes, which are currently not possible to obtain using fieldable benchtop NGS technologies. Developing this bioinformatic capability to work in conjunction with available technology would enable routine, inexpensive, rapid assessment of genetic markers correlated with not only clinically important phenotypes, but forensically important signatures as well. This will shorten the sample-to-forensic-answer time from days or hours (sequencing) potentially to minutes.

Build an objective "tool" for discriminating between natural and other outbreaks based on molecular evidence

We have mentioned several "tools" that use epidemiological data to determine whether an outbreak is mostly likely natural or unnatural (Dembek et al., 2007; Grunow & Finke, 2002; Radosavljevic & Belojevic, 2012). These utilize both quantitative and qualitative epidemiological data about an outbreak and require a *subjective interpretation* (e.g., see the numbered list above in the *Epidemiological factors to discern between deliberate, accidental, and natural outbreaks* section). We suggest the development of a *molecular epidemiology tool* that relies on *objective data* rather than *subjective interpretation*. For example, instead of epidemiological criteria such as "higher morbidity than expected" (Dembek et al., 2007) or "unusual disease symptoms" (Radosavljevic & Belojevic, 2012), both of which are contingent on subjective interpretation by a subject-matter expert, molecular epidemiology criteria such as "number of multiple cloning sites indicative of genetic engineering" or "nucleotide substitution rate as determined by SNP analysis" could be used in such a tool. While the bioinformatician and subject-matter expertise requirements would not be eliminated, we believe these objective, molecular epidemiologic approaches are less conditional on individual expert interpretation and will have less interpretational variance. Development of such a tool would require both a scoring system and a set of thresholds that need to be validated against molecular data, preferably obtained from real-life natural, accidental, and deliberate infectious disease outbreaks. The new "tool" would need to be refined and revalidated as thresholds are tested in actual field use and experience is gained from actual users and decision-makers.

Build more complete reference databases

In many instances, the biggest limitation is not the software or the sequencing technology, but the completeness and correctness of the reference database being used (see Figure 1). None of the methods discussed in the bioinformatics section above will perform well when trying to match a sequence read or assembly to a database that doesn't contain a labeled sequence corresponding to the source organism for the query sequence.

We cannot emphasize strongly enough the importance of complete, always-up-to-date reference databases against which to compare query samples (whether those be unassembled reads or draft/complete genome assemblies). A path to field-deployable capability must include some level of critical assessment and quality control of the existing publicly available data, as well as sequencing and cataloging new and emerging strains as they become available. Fortunately, advances in third-generation sequencing technology throughput and error correction enable finished *de novo* microbial genome assemblies with up to 99.999% accuracy (C.-S. Chin et al., 2013; Koren et al., 2012). Field-deployable capability would also include a plan for keeping reference databases consolidated and always up-to-date.

From a forensic attribution perspective, the utility of interpretation software will depend on a continuously updated database that would not only encompass pathogen genomes to detect presence of a species, but would also contain a catalog of point mutations or genes that account for drug resistance, virulence, and marks of genetic engineering. Furthermore, detection of genetic engineering is challenging due not only to the large number of cloning vectors that may closely resemble the organism from which they are derived (e.g. *E. coli*), but also because of the availability of numerous "scarless" genetic engineering procedures (Blank, Hensel, & Gerlach, 2011; Fehér et al., 2008; R. Liang & Liu, 2010; Sun, Wang, & Curtiss, 2008). Robust and high-resolution assessment of these genetic signatures using NGS increases the confidence with which once can make a statement about the likely release being deliberate, accidental, or natural. However, every day new mechanisms of antibiotic resistance are emerging (naturally and potentially artificially), new virulence factors are discovered, and genetic engineering vectors and methodology are improved. Keeping consolidated, always-up-to-date references for each of these is required for any effective deployed capability.

Furthermore, it is important to recognize that all relevant reference populations will not necessarily be available data in a database (Budowle, Schutzer, Breeze, Keim, & Morse, 2011). Methods that map genome sequence onto reference databases do so probabilistically, accounting for both the degree of similarity as well as the size of the reference database. Thus, the probability of an apparent "match" depends highly on the size of the reference database, which may include only a portion of the total genetic data available. The size of this portion varies globally, regionally and locally. Further investment needs to be made in world-wide sampling, sequencing, cataloging, and curating of local backgrounds such that endemicity comparisons can be made on-demand. These comparisons are critical in assessing the likelihood of a disease outbreak being deliberate, accidental, or natural.

Finally, much of what has been discussed here involves comparing a query sample to a known reference (Q-K comparison). It is feasible that there could be forensic value, or "value in the clutter" of a metagenomic query sample. As such, a field-deployable capability will have a plan in place for collecting, archiving, and properly tracking query sample metadata, such that the samples could be compared to other query samples (Q-Q comparisons). Although a bioweapon might not have been discovered in "suspect lab A" (regardless of whether a weapon was absent or if the sequencing depth was too shallow to pick it up), forensic value could be realized if the metagenomic profile of "suspect

lab A" matched the metagenomic profile of "suspect lab B" to the reasonable exclusion of other endemic samples. In such a way, different locations, personnel, and methods could be tagged with signatures for further, future investigation. As such, understanding endemic background and creation/maintenance of a query/background sample reference database is key in any field-deployed capability.

Develop data integration & decision analytics procedures

Finally, a field-deployed capability will have guidelines and standard operating procedures in place for: (1) integrating genetic forensic evidence with other forms of evidence (conventional genetic assays, epidemiological evidence, non-genetic evidence, classical forensics); and (2) integrating a combined and properly weighted forensic analysis with other intelligence for making a tactical, strategic, legal, or policy decision. Consider that the deliberate release of a biological threat is an exceedingly rare event, compared to natural outbreaks of disease, which may appear unusual. Even after registering an infectious disease outbreak as *unusual*, the likelihood that it is a *deliberate* attack is still very low. Thus to minimize economic, political, and human impact after an outbreak is registered as atypical, molecular data must be interpreted in the context of many other forms of evidence and intelligence.

NGS is one element of a microbial forensics capability (albeit a piece that has the potential to replace many other lower-resolution techniques). A discussion of data integration with respect to intelligence analysis is beyond the scope of this report, but a field-deployable capability must include consideration of how best to integrate heterogeneous data types, weighing each fragment of data inversely to the confidence it increases in one potential attribution to the exclusion of others.

Conclusions

We have: (1) critically reviewed NGS and bioinformatics technology as they might be applied to microbial forensics; (2) given examples of state-of-the-art approaches from public health; and (3) designed a path forward to future deployment capability for using these technologies in determining whether an outbreak is the result of a natural, accidental, or deliberate event.

Two conclusions were drawn at the 1998 NATO Advanced Research Workshop on Scientific and Technical Means of Distinguishing Between Natural and Other Outbreaks of Disease (Proceeding of the NATO Advanced Research Workshop, 2001). Regarding natural versus other outbreaks of disease:

> *It was recognized that once an outbreak is initiated its spread occurs naturally and consequently all outbreaks are "natural" whether arising from a natural occurrence or from a deliberate or accidental initiation of the outbreak. ... It was evident that there is no unique signature that distinguishes an outbreak from other causes from an outbreak resulting from a natural cause, as many natural outbreaks have unusual characteristics.*

We still agree that the first part of this statement is correct – after an outbreak has been initiated, its spread and life-cycle will progress "naturally". However, we have shown that new developments over the last decade have rendered the second part of this statement partially incorrect. We have shown through case studies in the public health literature and by reviewing the capabilities of current and future NGS and bioinformatics technologies that often there *is* a unique signature that can distinguish

between natural and other outbreaks of disease. However, it is still true that many "natural" outbreaks can have unusual characteristics, like the dangerous cocktail of toxins, virulence factors, and antibiotic-resistance genes present in the *E. coli* O104:H4 example discussed above.

A second conclusion drawn from the same NATO workshop (Proceeding of the NATO Advanced Research Workshop, 2001) states:

> *The investigation approach and techniques would be similar for all outbreaks, whether occurring naturally or unnaturally. In the end the interpretation of investigation results by experts would be the only method to determine the di*fference. ... *In the absence of a "smoking gun" it would be virtually impossible to provide proof that an outbreak was unnatural in origin.*

This statement, even in light of advances in NGS and bioinformatics technology, is still applicable. Investigations will be initiated often without prior knowledge of whether an outbreak is due to natural or other causes, and the techniques used will be similar. Consider that nature has been much more creative than any known terrorist or enemy combatant at recombining and mutating genes to create well-adapted microorganisms that are often pathogenic to humans. Combined with the fact that true deliberate releases of biological weapons are rare, this means that most emerging pathogens will be naturally occurring. Using NGS data alone is usually not enough to make the distinction between accidental or deliberate, which implies a notion of *intent*. The elements of intent and human motivation cannot be gleaned from NGS data alone in the absence of traditional epidemiological "detective work," no matter how sophisticated the bioinformatics analysis. We also agree that interpretation of the investigation results by subject matter experts is in the end the only viable method to determine the difference between deliberate, accidental, and natural outbreaks of infectious diseases.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. doi:10.1016/S00222836(05)80360-2

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B., & Allen, J. E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics (Oxford, England)*.

Bae, W. K., Lee, Y. K., Cho, M. S., Ma, S. K., Kim, S. W., Kim, N. H., & Choi, K. C. (2006). A case of hemolytic uremic syndrome caused by Escherichia coli O104:H4. *Yonsei Medical Journal*, *47*(3), 437–9.

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4), 333–337. doi:10.1038/nmeth.1935

Bazinet, A. L., & Cummings, M. P. (2012). A Comparative Evaluation of Sequence Classification Programs. *BMC Bioinformatics*, *13*(92), 1–13.

Berendzen, J., Bruno, W. J., Cohn, J. D., Hengartner, N. W., Kuske, R., Mcmahon, B. H., ... Xie, G.

(2012). Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. *BMC Research Notes*, *5*, 460.

Blank, K., Hensel, M., & Gerlach, R. G. (2011). Rapid and highly efficient method for scarless mutagenesis within the Salmonella enterica chromosome. *PloS one*, *6*(1), e15763. doi:10.1371/journal.pone.0015763

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., & Corbeil, J. (2012). Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biology*, *13*(12), R122. doi:10.1186/gb-2012-13-12-r122

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, İ., ... Korf,

I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.

Brady, A., & Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods*, *8*(5), 367. doi:10.1038/nmeth0511-367

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–6. doi:10.1038/nmeth.1358

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv*, *1203.4802*(v2), 1–18.

Bryant, J. M., Grogono, D. M., Greaves, D., Foweraker, J., Roddick, I., Inns, T., ... Floto,

R. A. (2013). Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. *Lancet*, *381*(9877), 1551–60. doi:10.1016/S0140-6736(13)60632-7

Budowle, B., Schutzer, S. E., Breeze, R. G., Keim, P. S., & Morse, S. A. (2011). *Microbial Forensics, Second Edition*. Elsevier.

Budowle, B., Schutzer, S. E., Morse, S. a, Martinez, K. F., Chakraborty, R., Marrone, B. L., ... Velsko, S. P. (2008). Criteria for validation of methods in microbial forensics. *Applied and Environmental Microbiology*, *74*(18), 5599–607. doi:10.1128/AEM.00966-08

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., ... Korlach,

J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *advance on*. doi:10.1038/nmeth.2474

Chin, C.-S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., ... Waldor, M. K. (2011). The origin of the Haitian cholera outbreak strain. *The New England journal of medicine*, *364*(1), 33–42. doi:10.1056/NEJMoa1012928

Dalton, R. (2001). Genetic sleuths rush to identify anthrax strains in mail attacks. *Nature*, *413*(6857), 657–8.

Daniel, R. (2005). The Metagenomics of Soil. *Nature Reviews. Microbiology*, *3*(6), 470–8.

Davenport, C. F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., ... Sprengel, F. (2012). Genometa–a fast and accurate classifier for short metagenomic shotgun reads. *PloS one*, *7*(8), e41224. doi:10.1371/journal.pone.0041224

DeLong, E. F. (2005). Microbial Community Genomics in the Ocean. *Nature Reviews. Microbiology*,

3(6), 459–69.

Delcher, A. L., Salzberg, S. L., & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. In A. D. Baxevanis (Ed.), *Current Protocols in Bioinformatics* (Vol. 10, pp. 10.3.1–10.3.18). John Wiley & Sons, Inc.

Dembek, Z. F., Kortepeter, M. G., & Pavlin, J. a. (2007). Discernment between deliberate and natural infectious disease outbreaks. *Epidemiology and Infection*, *135*(3), 353–71. doi:10.1017/S0950268806007011

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, *13*(August), 601–612. doi:10.1038/nrg3226

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., ... Paten, B. (2011). Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Research*, *21*(12), 2224–41.

Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, *30*(4), 295–6. doi:10.1038/nbt0412-295

Enright, M. C., Robinson, D. A., Randle, G., Feil, E. J., Grundmann, H., & Spratt, B. G. (2002). The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proceedings of the National Academy of Sciences of the United States of America*, *99*(11), 7687–92. doi:10.1073/pnas.122108599

Ervin, A., Hultgren, A., Rhyne, E., & Ward, K. (2010). Sensing Dispersal of Chemical and Biological Agents in Urban Environments. In *Wiley Handbook of Science and Technology for Homeland Security* (pp. 423–434).

Fehér, T., Karcagi, I., Gyorfy, Z., Umenhoffer, K., Csörgo, B., & Pósfai, G. (2008). Scarless engineering of the Escherichia coli genome. *Methods in molecular biology (Clifton, N.J.)*, *416*, 251–9. doi:10.1007/978-1-59745-321-9_16

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., ... Caporaso,

J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390–5. doi:10.1073/pnas.1215210110

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods*, *6*(11 Suppl), S6–S12. doi:10.1038/nmeth.1376

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., ... Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, *7*(6), 461–5.

Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., an der Heiden, M., ... Krause, G. (2011). Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. *The New England Journal of Medicine*, *365*(19), 1771–80. doi:10.1056/NEJMoa1106483

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., ... Tang,

P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England journal of medicine*, *364*(8), 730–9. doi:10.1056/NEJMoa1003176

Gault, G., Weill, F. X., Mariani-Kurkdjian, P., Jourdan-da Silva, N., King, L., Aldabe, B., .. . Rolland,

P. (2011). Outbreak of haemolytic uraemic syndrome and bloody diarrhoea due to *Escherichia coli* O104:H4, south-west France, June 2011. *Euro Surveillance*, *16*(26).

Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., & Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC bioinformatics*, *10*, 430. doi:10.1186/1471-2105-10-430

Gevers, D., Pop, M., Schloss, P. D., & Huttenhower, C. (2012). Bioinformatics for the Human Microbiome Project. *PLoS Computational Biology*, *8*(11), e1002779.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., ... Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1513–8. doi:10.1073/pnas.1017351108

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. doi:10.1186/gb-2010-11-8-r86

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a, Amit, I., ... Regev,

A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7).

Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., Fitzgerald, M., .. . Hanage, W. P. (2012). Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), 3065–70. doi:10.1073/pnas.1121491109

Grunow, R., & Finke, E.-J. (2002). A procedure for differentiating between the intentional release of biological warfare agents and natural outbreaks of disease: its use in analyzing the tularemia outbreak in Kosovo in 1999 and 2000. *Clinical microbiology and infection*, *8*(8), 510–21.

Haft, D. H., & Tovchigrechko, A. (2012). High-speed microbial community profiling. *Nature Methods*, *9*(8), 793–794. doi:10.1038/nmeth.2080

Hall, B. G., Cardenas, H., & Barlow, M. (2013). Using complete genome comparisons to identify sequences whose presence accurately predicts clinically important phenotypes. *PloS one*, *8*(7), e68901.

Hall, N. (2013). After the gold rush. *Genome biology*, *14*(5), 115.

Harris, S. R., Cartwright, E. J., Török, M. E., Holden, M. T., Brown, N. M., Ogilvy-Stuart,

A. L., ... Peacock, S. J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet infectious diseases*, *13*(2), 130–6. doi:10.1016/S1473-3099(12)70268-2

Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., ... Bentley, S. D. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, *327*(5964), 469–74. doi:10.1126/science.1182395

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., ... Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N.Y.)*, *331*(6016), 463–7. doi:10.1126/science.1200387

Human Microbiome Project Consortium. (2012). A framework for human microbiome research.

Nature, *486*(7402), 215–21. doi:10.1038/nature11209

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–86. doi:10.1101/gr.5969107

Inglesby, T. V. (1999). Anthrax: A possible case history. *Emerging infectious diseases*, *5*(4), 556–60.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, *44*(2), 226–32. doi:10.1038/ng.1028

Iqbal, Z., Turner, I., & McVean, G. (2012). High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics*. doi:10.1093/bioinformatics/bts673

Jansen, A., & Kielstein, J. T. (2011). The new face of enterohaemorrhagic Escherichia coli infections. *Euro Surveillance*, *16*(25).

Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., ... Pallen, M. J. (2013). Reply to Updating benchtop sequencing performance comparison. *Nature biotechnology*, *31*(4), 294–6.

Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nature Reviews. Microbiology*, *2*(2), 123–40.

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13770–3.

Kembel, S. W., Jones, E., Kline, J., Northcutt, D., Stenson, J., Womack, A. M., ... Green, J.

L. (2012). Architectural design influences the diversity and structure of the built environment microbiome. *The ISME journal*, *6*(8), 1469–79. doi:10.1038/ismej.2011.211

Khan, A. S., Levitt, A. M., & Sage, M. J. (2000). Biological and Chemical Terrorism: Strategic Plan for Preparedness and Response. Recommendations of the CDC Strategic Planning Workgroup. *Morbidity and Mortality Weekly Report*, *49*(RR04), 1–14.

Klevens, R. M., Morrison, M. A., Nadle, J., Petit, S., Gershman, K., Ray, S., ... Fridkin,

S. K. (2007). Invasive methicillin-resistant Staphylococcus aureus infections in the United States. *JAMA*, *298*(15), 1763–71. doi:10.1001/jama.298.15.1763

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., ... Adam M Phillippy. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, *30*(7), 693–700.

Krause,L.,Diaz,N.N.,Goesmann,A.,Kelley,S.,Nattkemper,T.W.,Rohwer,F.,... Stoye,

J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, *36*(7), 2230–9. doi:10.1093/nar/gkn038

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., & Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Reviews. Genetics*, *13*(1), 47–58.

Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., ... Peacock, S. J. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA

outbreak. *The New England Journal of Medicine*, *366*(24), 2267–75. doi:10.1056/NEJMoa1109910

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9.

Le, V. T. M., & Diep, B. A. (2013). Selected insights from application of whole-genome sequencing for outbreak investigations. *Current opinion in critical care*. doi:10.1097/MCC.0b013e3283636b8c

Lee, Z. M.-P., Bussema, C., & Schmidt, T. M. (2009). rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Research*, *37*(Database issue), D489–93.

Li,R.,Zhu,H.,Ruan,J.,Qian,W.,Fang,X.,Shi,Z.,... Wang,J.(2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–72. doi:10.1101/gr.097261.109

Liang, R., & Liu, J. (2010). Scarless and sequential gene modification in Pseudomonas using PCR product flanked by short homology regions. *BMC microbiology*, *10*, 209. doi:10.1186/1471-2180-10-209

Lipsitch, M., Finelli, L., Heffernan, R. T., Leung, G. M., & Redd, S. C. (2011). Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecurity and Bioterrorism: biodefense strategy, practice, and science*, *9*(2), 89–115.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., & Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, *12 Suppl 2*, S4. doi:10.1186/1471-2164-12-S2-S4

Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z., Quick, J., ... Pallen, M. J. (2013). A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4 Outbreak of Shiga-toxigenic Escherichia coli. *JAMA*, *309*(14), 1502. doi:10.1001/jama.2013.3231

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, *30*(5), 434–9. doi:10.1038/nbt.2198

MacDonald, N. J., Parks, D. H., & Beiko, R. G. (2012). Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Research*, *40*(14), e111. doi:10.1093/nar/gks335

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., ... Salzberg, S. L. (2013). GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms. *Bioinformatics (Oxford, England)*, *29*(14), 1718–25. doi:10.1093/bioinformatics/btt273

Mariani-Kurkdjian, P., Bingen, E., Gault, G., Jourdan-Da Silva, N., & Weill, F.-X. (2011). Escherichia coli O104:H4 south-west France, June 2011. *The Lancet infectious diseases*, *11*(10), 732–3. doi:10.1016/S1473-3099(11)70266-3

Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., ... Kyrpides, N. C. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research*, *40*(Database issue), D123–9. doi:10.1093/nar/gkr975

Markowitz,V.M.,Chen,I.-M.a,Palaniappan,K.,Chu,K.,Szeto,E.,Grechkin,Y.,... Kyrpides, N. C. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(Database issue), D115–22. doi:10.1093/nar/gkr1044

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72.

Mellmann, A., Bielaszewska, M., Köck, R., Friedrich, A. W., Fruth, A., Middendorf, B., ... Karch, H. (2008). Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic Escherichia coli. *Emerging infectious diseases*, *14*(8), 1287–90. doi:10.3201/eid1408.071082

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., ... Karch, H. (2011). Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PloS One*, *6*(7), e22751. doi:10.1371/journal.pone.0022751

Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., & Yampolskaya, O. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science (New York, N.Y.)*, *266*(5188), 1202–8.

Metzker, M. L. (2010). Sequencing technologies -the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. doi:10.1038/nrg2626

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., ... Edwards, R.

A. (2008). The metagenomics RAST server -a public resource for the automatic phylogenetic and functional analysis of metagenomes. BM*C* bioinformatics, 9, 386. doi:10.1186/1471-2105

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–27. doi:10.1016/j.ygeno.2010.03.001

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 1–12. doi:10.1093/nar/gks678

Narzisi, G., & Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS One*, *6*(4), e19175. doi:10.1371/journal.pone.0019175

National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington DC: National Academies Press.

Oinn,T.,Addis,M.,Ferris,J.,Marvin,D.,Senger,M.,Greenwood,M.,... Li,P.(2004).Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–54. doi:10.1093/bioinformatics/bth361

Pallen, M. J. (2013). Reply to Updating benchtop sequencing performance comparison. *Nature biotechnology*, *31*(4), 296. doi:10.1038/nbt.2531

Patil, K. R., Roune, L., & McHardy, A. C. (2012). The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PloS one*, *7*(6), e38581. doi:10.1371/journal.pone.0038581

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., & Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(33), 13272–7. doi:10.1073/pnas.1121464109

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*, *27*(13), i94–101. doi:10.1093/bioinformatics/btr216

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, *28*(11), 1420–8. doi:10.1093/bioinformatics/bts174

Proceeding of the NATO Advanced Research Workshop. (2001). *Scientific and Technical Means of Distinguishing Between Natural and Other Outbreaks of Disease*. (M. R. Dando, G.

S. Pearson, & B. Kris, Eds.) (35 ed.). Prague: Springer. Qin, J., Li, R., Raes, J., Arumugam, M.,

Burgdorf, K. S., Manichanh, C., ... Ehrlich, S.

D. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. doi:10.1038/nature08821

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... Gu,

Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC ...* , *13*(12), 341. doi:10.1186/14712164-13-341

Radosavljevic, V., & Belojevic, G. (2012). Unusual epidemic events: A new method of early orientation and differentiation between natural and deliberate epidemics. *Public Health*, *126*(1), 77–81. doi:10.1016/j.puhe.2011.11.006

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*, 369–94. doi:10.1146/annurev.micro.57.030502.090759

Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., ... Waldor, M. K. (2011). Origins of the E. coli strain causing an outbreak of hemolyticuremic syndrome in Germany. *The New England Journal of Medicine*, *365*(8), 709–17. doi:10.1056/NEJMoa1106920

Reimer, A. R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., ... Gerner-Smidt, P. (2011). Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerging Infectious Diseases*, *17*(11), 2113–21. doi:10.3201/eid1711.110794

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome biology*, *14*(6), 405.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome biology*, *14*(6), 405. doi:10.1186/gb-2013-14-6-405

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., ... Niemann, S. (2013). Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS medicine*, *10*(2), e1001387. doi:10.1371/journal.pmed.1001387

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M., ... Yang, R. (2011). Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *The New England Journal of Medicine*, *365*(8), 718–24.

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, *27*(1), 127–9. doi:10.1093/bioinformatics/btq619

Rotz, L. D., & Hughes, J. M. (2004). Advances in detecting and responding to threats from bioterrorism and emerging infectious disease. *Nature Medicine*, *10*(12 Suppl), S130–6. doi:10.1038/nm1152

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ... Frazier, M. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, *5*(3), e77. doi:10.1371/journal.pbio.0050077

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–92.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–4.

Sharma, V. K., Kumar, N., Prakash, T., & Taylor, T. D. (2012). Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. Plo*S* One, *7*(4), e34030. doi:10.1371/journal.pone.0034030

Sherwood, J., Park, M., Robben, P., Whitman, T., & Ellis, M. W. (2013). USA300 methicillinresistant Staphylococcus aureus emerging as a cause of bloodstream infections at military medical centers. *Infection control and hospital epidemiology*, *34*(4), 393–9. doi:10.1086/669866

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117–23. doi:10.1101/gr.089532.108

Sun, W., Wang, S., & Curtiss, R. (2008). Highly efficient method for introducing successive multiple scarless gene deletions and markerless gene insertions into the Yersinia pestis chromosome. *Applied and environmental microbiology*, *74*(13), 4241–5. doi:10.1128/AEM.00940-08

Sutton, V. (2005). Dual Purpose Bioterrorism Investigations in Law Enforcement and Public

Health Protection: How to Make them Work Consistent with the Rule of Law. *Houston Journal of Health Law & Policy*, *6*, 151–170.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics -a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. doi:10.1186/2042-5783-2-3

Treadwell, T. A., Koo, D., Kuker, K., & Khan, A. S. (2003). Epidemiologic clues to bioterrorism. *Public Health Reports*, *118*(2), 92–98.

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., ... Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology*, *14*(1), R2.

Tremlett, G, Pidd, H. (2011, may). Germany Admits Spanish Cucumbers are not to Blame for E. Coli Outbreak.

Turner, S. (2013). Bioinformatics Training Resources. *Figshare*. doi:10.6084/m9.figshare.773083

Vezzi, F., Narzisi, G., & Mishra, B. (2012). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, *7*(12), e52210. doi:10.1371/journal.pone.0052210

Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., ... Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases*, *13*(2), 137–46. doi:10.1016/S1473-3099(12)70277-3

Wang, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *Journal of Computational Biology*, *19*(2), 241–9. doi:10.1089/cmb.2011.0276

Wetterstrand, K. S. (2013). DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program.

Wilkening, J., Wilke, A., Desai, N., & Meyer, F. (2009). Using clouds for metagenomics: A case study. *2009 IEEE International Conference on Cluster Computing and Workshops*, 1–6. doi:10.1109/CLUSTR.2009.5289187

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: read length matters. *Applied and environmental microbiology*, *74*(5), 1453–63. doi:10.1128/AEM.02181-07

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, *6*(2), e1000667.

World Health Organization. (2008). *International Health Regulations (2005)* (2nd ed.). Geneva: World Health Organiziation.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–9. doi:10.1101/gr.074492.107

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS One*, *6*(3), e17915.

Zinderman, C. E., Conner, B., Malakooti, M. A., LaMar, J. E., Armstrong, A., & Bohnker, B.

K. (2004). Community-acquired methicillin-resistant Staphylococcus aureus among military recruits. *Emerging infectious diseases*, *10*(5), 941–4. doi:10.3201/eid1005.030604