Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection

Report from the Winter Mid-Atlantic Microbiome Meet-up, College Park, MD January 10th, 2018

Jacquelyn S. Meisel[1], Daniel J. Nasko[1], Brian Brubach[1], Victoria Cepeda Espinoza[1], Jessica Chopyk[2], Héctor Corrada-Bravo[1], Marcus Fedarko[1], Jay Ghurye[1], Kiran Javkar[1], Nathan D. Olson[1,3], Nidhi Shah[1], Sarah M. Allard[2], Adam L. Bazinet[4], Nicholas H. Bergman[4], Alexis Brown[5], J Gregory Caporaso[6], Sean Conlan[7], Jocelyne DiRuggiero[8], Samuel P. Forry[3], Nur A. Hasan[1,9], Jason Kralj[3], Paul M. Luethy[10], Donald K. Milton[11], Brian D. Ondov[1,7], Sarah Preheim[12], Shashikala Ratnayake[4], Stephanie M. Rogers[13], M.J. Rosovitz[4], Eric G. Sakowski[12], Nils Oliver Schliebs[14], Daniel D. Sommer[4], Krista L. Ternus[15], Gherman Uritskiy[8], Sean X. Zhang[16], Mihai Pop[1], Todd J. Treangen[1*▽]

[1] Center for Bioinformatics and Computational Biology, University of Maryland College Park, College Park, MD, USA

[2] School of Public Health, University of Maryland College Park, College Park, MD, USA

[3] Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

[4] National Biodefense Analysis and Countermeasures Center, Frederick, MD, USA

[5] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

[6] The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

[7] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

[8] Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[9] CosmosID Inc., Rockville, MD, USA

[10] Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA

[11] Maryland Institute for Applied Environmental Health, School of Public Health, University of Maryland College Park, College Park MD, USA

[12] Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, USA

[13] B.Next, In-Q-Tel, Inc., Arlington, VA, USA

[14] Department of Computer Science, University of Tübingen, Tübingen, Germany

[15] Signature Science, LLC, Arlington, VA, USA

[16] Division of Medical Microbiology, Department of Pathology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

[▽] Current address: Department of Computer Science, Rice University, Houston, TX, USA

* Address correspondence to:

Todd Treangen

Department of Computer Science – MS-132

Rice University

P.O. Box 1892

Houston, TX 77005-1892

713-348-4724

treangen@rice.edu

Abstract

The Mid-Atlantic Microbiome Meet-up (M^3) organization brings together academic, government, and industry groups to share ideas and develop best practices for microbiome research. In January of 2018, M^3 held its fourth meeting, which focused on recent advances in biodefense, specifically those relating to infectious disease, and the use of metagenomic methods for pathogen detection. Presentations highlighted the utility of next-generation sequencing technologies for identifying and tracking microbial community members across space and time. However, they also stressed the current limitations of genomic approaches for biodefense, including insufficient sensitivity to detect low abundance pathogens and the inability to quantify viable organisms. Participants discussed ways in which the community can improve software usability and shared new computational tools for metagenomic processing, assembly, annotation, and visualization. Looking to the future, they identified the need for better bioinformatics toolkits for longitudinal analyses, improved sample processing approaches for characterizing viruses and fungi, and more consistent maintenance of database resources. Finally, they addressed the necessity of improving data standards to incentivize data sharing. Here, we summarize presentations and discussions from the meeting, identifying areas where microbiome analyses have improved our ability to detect and manage biological threats and infectious disease, as well as gaps of knowledge in the field that require future funding and focus.

Keywords

Microbiome; Metagenomics; Bioinformatics; Biodefense; Bio-threats; Pathogen Detection; Longitudinal Analysis

Abbreviations

CBCB – Center for Bioinformatics and Computational Biology

CONSERVE – Center of Excellence at the Nexus of Sustainable Water Reuse, Food, and Health

CRISPR – clustered regularly interspaced short palindromic repeats

CPU – central processing unit

FPGA – field programmable gate array

GPU – graphics processing unit

IQT – In-Q-Tel, Inc.

JHU – Johns Hopkins University

M^3 – Mid-Atlantic Microbiome Meet-up

NAU – Northern Arizona University

NBACC – National Biodefense Analysis and Countermeasures Center

NGS – next-generation sequencing

NHGRI – National Human Genome Research Institute

NIH – National Institutes of Health

NIST– National Institute of Standards and Technology

RAM – random access memory

SPH – School of Public Health

UMD – University of Maryland

Introduction

Strong public health and biodefense research are essential for the prevention, detection, and management of biological threats and infectious disease. Over the last century, the focus of biodefense research has shifted in response to modern advances in biotechnology. Specifically, a biological revolution is underway, generating promising new gene editing and synthetic biology technologies that may transform modern medicine, but also present a threat to public health if misappropriated [1]. As biotechnology becomes increasingly globalized, it is important that we establish new strategies and tools for infectious disease detection and surveillance that will help us protect against bioterrorism and manage disease outbreaks.

Rapid advances in next generation sequencing (NGS) technologies have helped advance biodefense research by enabling the development of new methods for identifying and characterizing pathogens. Amplification and sequencing of the 16S rRNA gene allows for high-throughput detection of prokaryotic communities, while shotgun metagenomic sequencing approaches capture the composition and functional potential of multi-domain populations. Metagenomic analyses used for pathogen detection and identification are often time sensitive. The results help inform high-stakes decision-making, such as choosing an appropriate medical treatment, deciding if a food product should be recalled due to contamination, or determining if an area should be shut down due to a suspected act of bioterrorism. In addition, geospatial and temporal metagenomic analyses are essential for tracking the dynamic responses of microbial populations to changes in environmental or human health. However, improvements in precision, sensitivity, speed, cost, and accuracy of NGS and downstream analyses are necessary for effective utilization in biodefense research [2-6].

On January 10, 2018, the Mid-Atlantic Microbiome Meet-up (M^3) organization held a conference aimed at understanding how the biodefense and pathogen detection fields are transformed by new biological and computational technologies. While biodefense was broadly discussed, the participants focused primarily on emerging infectious disease applications. The meeting took place in the STAMP Student Union on the University of Maryland campus in College Park. The M^3 consortium brings together microbiome researchers from different sectors to discuss challenges, develop standards and best practices, and help connect data generators with data analysts [7]. The M^3 community is constantly growing and, as of this publication, has 140 members from over 25 different institutions. The conference was attended by 67 participants from academia, government and industry **(Fig 1)**, with expertise in areas such as biodefense, computer science, genomics, microbiology, and public health. There were two talks given by invited speakers, 15 oral presentations selected from submitted abstracts, and several posters displayed at the meeting **(Supplementary Table 1)** [8]. Additionally, there were three interactive break-out sessions to address challenges of the field and encourage networking **(Supplementary Table 2)**. The event was sponsored in part by CosmosID, Inc. but they did participate in the organization of the event, nor in the selection of speakers and topics being discussed.

The tone for the meeting was set by the keynote address presented by Dr. Tara O'Toole, Executive Vice President of the non-profit strategic investor In-Q-Tel, Inc.. Pointing to the problems in detection, containment, and treatment during the recent H1N9 pandemic and Ebola epidemic, Dr. O'Toole shared that current progress in the field is disappointing because biodefense is not a priority for any single government agency, funding support is irregular, and epidemics are becoming more common. Increasing international competition for biotechnology advancements and leadership make it even more important to stimulate progress.

Dr. O'Toole outlined several keys to innovation and policy, which were echoed by the presentations and discussions throughout the remainder of the meeting, including (1) the willingness to think anew, (2) development of new tools and instruments, (3) implementation of a technology-focused biodefense strategy, (4) delivery of near real-time situational awareness for existing epidemics by leveraging modern data analytics and networked communications, and (5) establishment of rich human networks and cross-sector partnerships between government agencies, the private sector, and academia.

Key Conclusions

We start by highlighting the key conclusions and recommendations identified by the participants in the meeting:

1. Sequencing-based assays frequently face challenges related to limits of detection and technical biases, and culturing or other enrichment strategies remain necessary in many applications. The accurate quantification of viable organisms or metabolic activity within complex metagenomic samples remains an open challenge that is unlikely to be solved through sequencing alone.
2. Current sample processing approaches tend to exclude viral and fungal/eukaryotic components of microbial communities. In the case of viruses, this problem is compounded by poor taxonomies and database resources.
3. Analytical approaches, community standards, and software for temporal data analysis have lagged behind rapidly increased generation of such data.
4. Robust bioinformatics tools are critical for future progress. These tools must be developed to better match the needs of end-users and must be subject to critical validation.

5. Data standards are essential for ensuring the quality and usefulness of shared datasets, but overly onerous reporting requirements discourage sharing. In cases where privacy is a concern, we must also develop solutions that allow for secure storage and processing of sensitive data.

These key recommendations are summarized in **Table 1** and more extensively discussed below.

1. Sequencing-based assays frequently lack sensitivity

While the biodefense community has benefited from high-throughput sequencing strategies, these methods are not always as sensitive as required. In some cases, culturing is still the most reliable method for detecting pathogens because standard sequencing pipelines are not always available and achieving required sequencing depths may be cost-prohibitive. Dr. Sarah Allard (UMD SPH) shared her work from CONSERVE (a Center of Excellence at the Nexus of Sustainable Water Reuse, Food, and Health), whose mission is to enable the safe use of non-traditional irrigation water sources on food crops [9]. Dr. Allard used both culture-based and sequence-based methods to detect foodborne pathogens in water samples. She concluded that culture-based techniques are currently the most sensitive pathogen detection strategies and that sequencing analysis sensitivity and stringency vary strongly by method.

From a public health perspective, quantification of viable organisms contributing to disease is essential, but cannot be achieved with metagenomic analysis alone. Culturing and other approaches are important for gaining insight into the metabolic activity of the microbes in a community [10]. Additionally, researchers must often make a trade-off between the sensitivity of their detection methods and the computational costs of analyzing increasingly deep sequencing

datasets. Even partial culturing of select organisms or samples can help shift this trade-off. As commented during a breakout session, “you can’t always sequence your way out of it.”

2. Few studies look beyond bacterial pathogens

Shotgun metagenomics and a decrease in the cost of DNA sequencing have enabled researchers to analyze the genetic potential of microorganisms directly from an environmental sample. However, the majority of microbiome and metagenome studies focus only on the prokaryotic component of the community, while few have explored the roles of fungi or viruses in these microbial communities. This is due, in large part, to limitations in resources, laboratory procedures and, in the case of viruses, the lack of a universally distributed marker gene. Additional barriers to mycobiome and virome studies, include the ability to obtain sufficient material from low biomass environments, high levels of host contamination, incomplete databases, and a lack of available wet lab protocols and computational analysis pipelines. At the meeting, it was noted that central repositories for shared protocols do exist (e.g. protocols.io [11]) and a concerted effort in viral protocol sharing has been made by the Gordon and Betty Moore Foundation, which funds VERVE Net [12]. Proposed goals to address other barriers included providing financial and/or publication incentives for database curation and maintenance and focusing work on gene function identification. Since the NCBI SRA already contains many metagenomic sequencing datasets, it may be worthwhile to identify novel fungal and viral genomes from existing datasets to optimize data usage, as this approach has been employed in previous studies of environmental viruses [13].

Despite the aforementioned barriers to fungal and viral metagenomics, additional research in this area can significantly contribute to biodefense. One such important topic is the spread of

viral pathogens. Invited seminar speaker Dr. Don Milton (UMD SPH) presented his work on the transmission of the influenza virus in college dormitories [14]. The Centers for Disease Control and Prevention (CDC) suggests that human influenza transmission mainly occurs by droplets made when people with flu cough, sneeze, or talk. However, Dr. Milton explained that dueling reviews have disputed the importance of airborne transmission [15-20]. He presented NGS data showing that exhaled breath of symptomatic influenza cases contains infectious virus in fine particles, suggesting that aerosol exposures are likely an important mode of transmission.

3. Tracking microbial communities across time and topography

Temporal and biogeographic sequencing studies provide increased resolution of microbial community shifts. In the context of biodefense, this is important for detecting and containing outbreaks. Additionally, these studies provide insight into environmental changes, which may contribute to epidemics by causing shifts in disease vectors and/or spurring human migration to new regions or densely-populated urban areas. Several presentations at the meeting shared spatiotemporal microbiome analyses of different environments. Dr. Sean Conlan (NIH, NHGRI) presented his work using metagenomics to study outbreaks of nosocomial infections and identified the transfer of plasmids from patients to the hospital environment [21,22]. Gherman Uritskiy (JHU) and Dr. Sarah Preheim (JHU) used a combination of marker gene and metagenomics approaches to characterize changes in environmental microbiomes in response to perturbations. Uritskiy studied halite endoliths from the Atacama Desert in Chile over several years and showed how they were significantly impacted by rainstorms. Dr. Preheim compared a biogeochemical model to microbial communities' changes in a lake over the spring and summer to reveal the influence of energy availability on microbial population dynamics.

While time-series datasets provide valuable information, they are much more difficult to analyze with current statistical methods and models than cross-sectional sampling strategies [23,24]. Among other reasons, this is because it is difficult to identify the optimal sampling frequency, the compositional nature of microbiome data frequently violates assumptions of statistical methods, and commonly available software tools are often insufficient for required complex comparisons. Addressing this, Dr. J Gregory Caporaso (NAU) presented QIIME 2 (https://qiime2.org) and shared his team's QIIME 2 plugin, q2-longitudinal, which incorporates multiple methods for characterizing longitudinal and paired-sample marker gene datasets [25].

4. Development and application of metagenomic analysis tools is critical for progress

Computational methods required for metagenomic analyses include taxonomic abundance profiling, taxonomic sequence classification and annotation, functional characterization, and metagenomic assembly. Many of the presentations at the meeting shared new and/or improved tools for different aspects of microbiome studies. Victoria Cepeda (UMD) described how her tool, MetaCompass, uses reference genomes to guide metagenome assembly [26] and Gherman Uritskiy (JHU) presented his pipeline, metaWRAP, for the pre-processing and binning of metagenomes [27]. Furthermore, Brian Ondov (UMD, NIH, NHGRI) shared his implementation of the MinHash containment estimation algorithm to screen metagenomes for the presence of genomes and plasmids [28]. Data visualization is important for accurately interpreting microbiome data analyses and Dr. Héctor Corrada-Bravo (UMD) demonstrated how to use his lab's tool, Metaviz [29], for interactive statistical analysis of metagenomes.

Conventional metagenomic analyses often reflect the most abundant elements from a complex sample and cannot detect rare elements with confidence. Dr. Nicholas Bergman (NBACC)

shared a more sensitive single cell metagenomics approach that allows for increased detection of all elements of a community sample. Dr. Bergman's talk also emphasized the necessity of improving sensitivity, preventing contamination, eliminating biases, and increasing efficiency for sequencing-based techniques.

Bioinformatics tools should better match the needs of end-users

Many discussions at the meetings focused on how the field can optimize tool utility. It was agreed that scientists should always carefully evaluate the strengths and weakness of available methods, either via existing "bake-off" studies or through the available documentation, to ensure they are using the best tools to address their specific problem. Tool developers should disclose the limits of their methods and advise on the types of data their software is best suited to analyze. Developers should also work towards producing software that is easy to download and install, providing comprehensive documentation for their tools, and ensuring open-access for the academic community. As a community, we should encourage that publications list not only cases and data types where methods perform best, but also where they under-perform or even fail. Additional studies, like the Critical Assessment of Metagenome Interpretation (CAMI) [30,31], Microbiome Quality Control project [32], or challenges run under the aegis of PrecisionFDA [33] should be conducted to help characterize strengths and weaknesses of different approaches and evaluate their impact on data analysis and interpretation.

Some meeting attendees are currently contributing to these goals. Dr. Nathan Olson (UMD, NIST) presented his evaluation of different 16S rRNA marker gene survey bioinformatic pipelines using mixture samples. Additionally, Dr. Daniel Nasko (UMD) characterized how genomic database growth affects study findings, showing that different versions of the RefSeq

297 database strongly influenced species-level taxonomic classifications from metagenomic
298 samples [34]. Because the version of software and databases used can significantly affect
299 findings, this information should be reported more consistently in the literature. Furthermore, we
300 should consider strategies to preserve previous software and database versions to enable
301 future replication of analyses.
302
303 *Bioinformatics tools must better navigate the trade-off between speed and accuracy*
304
305 Metagenomic analysis methods vary in central processing unit (CPU) time, memory, and disk
306 resource usage and this is not always clearly reported in software publications. Additionally,
307 method scalability relative to size or type of input data also varies considerably. Optimizing
308 speed and accuracy is especially important for biodefense applications. For instance,
309 improvements in NGS analysis allowing for collection and analysis of samples in a clinically
310 relevant time frame can help effectively track hospital outbreaks and prevent the spread of
311 infection [35]. Furthermore, confidence in the accuracy of these analyses is required to execute
312 appropriate plans of action and prevent panic. Recently, findings of *Bacillus* strains on the
313 International Space Station that were genomically similar to pathogenic *B. anthracis* required
314 more detailed characterization to ensure that their presence was not a concern for the health of
315 the crew [36-38]. *B. anthracis* was also initially reported to be found in the NYC subway system,
316 along with *Yersina pestis*, the pathogen responsible for the plague [39]. After public attention
317 prompted further analysis, the authors found no evidence that these organisms were present
318 and found no evidence of pathogenicity [40,41], again highlighting the importance of careful
319 evaluation and interpretation of results, especially those with severe public health
320 consequences.
321

Many different strategies for speeding up analyses were discussed at the meeting, including hardware, software, and algorithm choice. Some hardware considerations for speed of analyses include balancing CPUs with co-processors such as graphics processing units (GPUs) or field programmable gate arrays (FPGAs), server configuration in terms of amount of random access memory (RAM), or disk storage type and speed. Programs and algorithms vary in accuracy as well as ease of parallelization. Often a slower yet parallelizable algorithm is preferred to one that is not parallelizable. If a program supports parallelism, consideration should be given to the type of hardware required. For example, some available options include large multicore servers for multithreaded applications, cluster nodes for distribution of compute jobs, or cloud computing solutions. Other strategies might involve analyzing only a subset of the data or using a smaller, application-specific reference database.

Finally, strategies discussed for speeding up time-critical analyses included employing a multi-tiered approach (e.g., a quick first pass followed by more detailed analyses [42]), and considering the suitability of various sequencing platforms for certain applications. Interventions or optimizations were discussed with regard to their impact on analysis accuracy and interpretation of results. Preferred solutions are the ones that provide both the desired speed and accuracy, though more often than not there is a trade-off between the two. The optimal balance also depends on the use case. Assessment and validation methods are required to characterize a method's speed and accuracy. It will be up to the subject matter experts to determine the desired accuracy level for each case and the extent to which they can sacrifice accuracy for speed.

5. Data needs to be moved out of private silos and into public repositories

Data sharing is continually a challenge that gets raised within the biological community, especially as DNA/RNA sequencing becomes more ubiquitous and tangible outside of core facilities [43]. This challenge is prevalent across multiple scientific disciplines, and was recently highlighted by the National Research Council as a priority for microbial forensics [44]. There are numerous reasons data are not being shared, including the need to protect personally identifiable information or intellectual property rights prior to publication, and the lack of sufficient infrastructure or manpower to upload at scale. However, leveraging this diversity and breadth of data will be important for an effective biodefense capacity, as well as other bioscience applications like healthcare, pharmaceuticals, agriculture, and industry. In order to incentivize data sharing, we need to evaluate and improve publicly available resources for storing and processing data.

Inherent altruism or obligation to share data should be met with as little friction as possible, and we need to incentivize openness. One incentive is academic credit through authorship on publications, though this will require combined efforts of researchers, journal editors, and funding agencies to better define what contributions constitute data authorship and what responsibilities data authors have [45,46]. Another potential incentive is the availability of free software for data analysis, and meeting participants debated the desirability and sustainability of service-based options (e.g. MG-RAST [47]) compared to user-installable software options (e.g. QIIME [48], mothur [49]). At the meeting, Dr. Nur A. Hasan (CosmosID, Inc.) highlighted the cloud-based metagenome tools and databases his company has to offer. There are also strong movements toward software sharing, such as the Astrophysics Source Code Library [50] and the Materials Resource Registry at NIST [51].

It is expected that some quality standard is needed to maintain useable, open repositories. Where that standard is set can affect how much data is shared. For example, a high bar may

ensure high quality sequences and comprehensive metadata but minimize sharing, while a lower quality bar will more likely move data out of silos. The solution may be a combination of repositories with varying standards, or a single repository which allows for varying degrees of annotation completeness and allows the user to modify searches based on that feature. It is important to note that a single repository may be difficult to reliably curate and manage at scale. Another option is distributed but federated systems, like used by the US Virtual Astronomical Observatory [52]. Groups like the Genomic Standards Consortium [53,54] are working towards improving data quality by supporting projects such as Minimum Information about any Sequence (MIxS) [55], which establishes standards for describing genomic data and provides checklists to help with annotation. We need to build a community consensus on how much metadata is required to make reporting less onerous for data providers but ensure data usability by others in the field.

Incentivizing open data sharing should not be the only solution, as some sensitive data cannot be openly shared due to privacy regulations (e.g. human genomes and Health Insurance Portability and Accountability Act regulations). Other sectors, such as the financial industry, have long been working on solutions to enable storage, transit, and operations of protected data. These solutions include software-based approaches (e.g., homomorphic encryption, Yao's protocol, secure fault-tolerant protocols, oblivious transfer) and hardware-based approaches (e.g. AES full disk encryption for data storage, Intel® Software Guard Extension for secure operations). Dr. Stephanie Rogers presented the GEMStone 2.0 project from B. Next, an IQT Lab, called SIG-DB, which explores homomorphic encryption and Intel Software Guard Extension (SGX) to securely search genomic databases [56]. Early results of applying these solutions to biological data are promising and should be explored more fully.

Conclusions

Overall, this meeting successfully brought together scientists from academia, government and industry to present their research and discuss how high-throughput genomics methods have stimulated interest and progress in biodefense and pathogen detection. Notably, meeting participants used NGS tools to identify the transfer of microbes from patients to their hospital environments, track the transmission of influenza in a community living space, study environmental shifts over time, and evaluate the safety of using non-traditional water sources on food crops. These studies, and others, have been partly driven by cheaper, more reliable sequencing technologies and improvements in computational analysis tools. Open-source software for sequence processing and quality control, taxonomic annotation, metagenomic assembly and binning, and data visualization have been essential for growth. Continued development of these resources will result in significant scientific advances.

Despite this progress, there are several limitations to using NGS approaches for biodefense problems. First and foremost, sequencing methods are unable to accurately quantify viable organisms from metagenomic samples, which is essential for identifying potential threats to public health. Beyond that, applications for which NGS approaches are well-suited still present many challenges. Although sequencing costs are steadily declining, it remains expensive to process, computationally analyze, and store the increasingly large datasets that are generated. Confident detection of infectious, but potentially rare pathogens in a community often requires very deep sequencing and scientists must make the appropriate speed, cost, and accuracy trade-offs to best answer their research questions. In many cases, sequencing experiments may need to be complemented with culturing, enrichment, or other targeted approaches. Because of these limitations, and others, researchers must be extremely careful when interpreting data to identify biothreats; reporting false positives without critical validation can have significant fiscal and public health consquences. Developing the capacity to identify not only when a potential

pathogen is present, but also at what levels it is actively contributing to infectious disease will greatly improve our response to biothreats. Another area that requires further investigation is the detection of antimicrobial resistance. While only briefly highlighted in the meeting talks about influenza and noscomial tracing, antimicrobial resistance poses a significant threat to public health and biodefense. Current metagenomic sequencing methods allow us to identify antimicrobial resistance genes from different environments, however these techniques cannot determine whether these genes are actively being expressed and are currently not practical for wide-spread adoption in clinical settings [57].

To date, few microbiome studies have focused on viral and fungal/eukaryotic organisms, despite their potentially important community interactions and roles in pathogenesis. In order to generate relevant virome and mycobiome datasets, we must improve sample processing techniques and dedicate resources to effectively curate and maintain publicly available databases. We also need to develop advanced statistical toolkits for analyzing longitudinal studies. In general, tool developers should focus on creating user-friendly, adaptable resources, with comprehensive documentation and clear descriptions of default settings and optional parameters. These tools must be critically evaluated for their appropriate use-cases; however, when looking for emerging threats, it will be necessary to develop validation approaches that do not require the use of gold standards.

In order to encourage additional growth, the greater scientific community should invest in expanding and enforcing clear standards for genomic datasets. If set appropriately, these standards will help incentivize data sharing and improve the quality and usability of public repositories. Additional focus should be on strengthening best practices and solutions for handling sensitive datasets that are subject to privacy regulations. Moving forward, active

conversations between researchers and policy makers will be essential to expand and implement these ideas in biodefense.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Not applicable

Competing interests

The conference was partly supported through funding from CosmosID, Inc.. The sponsor did not participate in the organization of the event or the selection of speakers and topics.

Funding

The meeting was supported in part by the Center for Health-related Informatics and Bioimaging, a Center organized under the MPowering the State Partnership between the University of

Maryland Baltimore and College Park campuses. JSM, BB, VCE, MF, JG, KJ, NS, and MP were supported in part by grants to MP, including grant R01-AI-100947 from the NIH and grant IIS-1513615 from the NSF. DJN and TJT were supported in part by the FunGCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. HC was supported by the NIH, R01 grant GM114267. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government. The contributions of ALB, NHB, MJR, DDS and SR were funded under Contract No. HSHQDC-15-C-00064 awarded by the Department of Homeland Security (DHS) Science and Technology Directorate (S&T) for the operation and management of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the DHS or S&T. In no event shall DHS, NBACC, S&T or Battelle National Biodefense Institute have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. DHS does not endorse any products or commercial services mentioned in this publication. JGC was supported in part by National Cancer Institute of the National Institutes of Health under the awards for the Partnership of Native American Cancer Prevention U54CA143924 (UACC) and U54CA143925 (NAU), and by the National Science Foundation award 1565100. SC was supported by NIH Intramural Research. JD and GU were supported in part by the NSF, grant DEB1556574 to JD. BDO was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health and utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

Authors' contributions

TT and MP organized the meeting. JK, JSM and DN reviewed abstracts, and SMR, NHB, and JD selected abstracts for the sessions they chaired. All listed authors contributed to the writing of the report or met at least one of the following requirements: they (1) gave an oral presentation, (2) presented a poster, and/or (3) attended and contributed to the interactive break-out sessions.

Acknowledgements

We would like to thank all those who helped make this meeting a success, especially Barbara Lewis (UMD) who organized the administrative aspects of the program and CosmosID, Inc. for funding the event. We would also like to thank Dr. Jayne Morrow and Dr. Robert Hanisch for providing helpful feedback on the manuscript.

Figure Legends

Figure 1. Different sectors and institutions represented at the January 2018 M^3 Meetup

527

Research Gaps	Current Limitations	Community Goals
Tracking microbial communities across time and topography (Key Conclusions 1 & 3) Importance: Studies incorporating temporal and/or spatial sampling allow us to detect important shifts in community dynamics Application example: Detecting the spread of infection in a hospital or of a pathogen contaminating crops and spreading food-borne illness	• Sequencing strategies are not able to quantify viable organisms (which is essential for biodefense applications) • Lack of well-established statistical approaches for exploring longitudinal microbiome data • Increased sample size makes these studies more expensive and harder to obtain sufficient statistical power for all subjects/timepoints/regions	• Collection, sequencing, and sharing of more time-series datasets • Development of statistical methods and tools to help analyze longitudinal and/or geospatial microbiome datasets
Looking beyond bacterial pathogens (Key Conclusion 2) Importance: Viral and fungal components of the microbiome are often under-explored, despite their potential implications in biodefense Application example: Better understanding the transmission of infectious viruses, like influenza	• Lack of a universally distributed marker gene (viruses) • Difficult to obtain sufficient material from low biomass environments • High levels of host contamination • Incomplete databases	• More consistent database curation and maintenance (potentially incentivized financially or with publications) • Improved gene function identification
Development and application of metagenomic analysis tools (Key Conclusion 4) Importance: Computational tools need to be developed to help improve the utility of high-throughput sequencing strategies for biodefense problems Application example: Improved metagenome assembly methods could better delineate between different strains of a pathogen in samples	• Tools for metagenome pre-processing, assembly, and binning are not always sensitive or fast enough for detection of pathogens in a sample • As sequencing technologies advance, we need new tools to handle output from long and short read technologies, as well as single cell metagenomics approaches	• Easy to install, open-access software with comprehensive documentation detailing best and worst use cases • Defined metrics for critical assessment and validation of existing tools • Software and database versions should be more consistently reported in the literature and preserved for future replication of analyses

528

Navigating the trade-off between speed and accuracy (Key Conclusion 4) Importance: Metagenomic analysis used for pathogen detection and identification are time-sensitive Application example: Deciding if a food product should be recalled due to contamination	• Current algorithms vary in speed and accuracy (often sacrificing one for the other) • Large datasets, error-prone heuristics, and coarse resolution of k-mer based methods present challenges	• Better documentation of available tools to help users optimize their software choice based on their available resources • Improvements in sequencing technologies and tools/algorithms to improve both speed and accuracy
Storing and sharing data (Key Conclusion 5) Importance: Access to publicly available datasets will help in verification of results and advance of scientific knowledge. Scientists need to be encouraged to move their data out of private silos and into shared databases	• Not all data can be shared because it is important to protect personally identifiable information or intellectual property rights • Lack of sufficient infrastructure or manpower to upload or store datasets at scale	• Defined quality standard to maintain usable, open repositories • Improved ways for secure interrogation of genomic datasets that cannot be openly shared due to privacy regulations

529

530 **Table 1.** Outline of current research gaps and future goals discussed at the January 2018 M^3

531 Meeting

532

533

534

	Speaker	Title
Keynote	Tara O'Toole In-Q-Tel, Inc.	Bioterror, and Biodefense, 2018
Invited Seminar	Don Milton UMD SPH	College Dorms as a Laboratory for Studying Respiratory Infection
Data-driven session Chair: Stephanie Rogers B.Next (In-Q-Tel, Inc.)	Stephanie Rogers B.Next (In-Q-Tel, Inc.)	Bridging technology, venture, and national security
	Daniel Nasko UMD CBCB	Tragedy of the commons: RefSeq database growth influences the accuracy and sensitivity of species identification from metagenomic samples
	Sarah Allard UMD SPH	Comparison of sequencing and culture-based methods for the detection of foodborne pathogens in non-traditional irrigation water in the Mid-Atlantic United States: A CONSERVE study
	Sean Conlan NHGRI, NIH	Tracking Antibiotic Resistance Across Space and Time
	Greg Caporaso NAU, NCI, NIH	Longitudinal analysis of microbiomes
Methods-driven session Chair: Nicholas Bergman NBACC	Nicholas Bergman NBACC	Metagenomics in bioforensics
	Brian Ondov UMD, NHGRI, NIH	Mash Screen: Fast sequence containment estimation using MinHash
	Nathan D. Olson UMD CBCB and NIST	A Sample Mixture Experiment to Assess 16S rRNA Metagenomic Methods
	Victoria Cepeda UMD CBCB	MetaCompass: Reference-guided Assembly of Metagenomes
	Héctor Corrada-Bravo UMD CBCB	Metaviz: Interactive Statistical and Visual Analysis of Human Microbiome Project Data

Ecology-driven session Chair: Jocelyne DiRuggerio JHU	Gherman Uritskiy JHU	Dynamic Response of Atacama Desert Extremophiles to Weather Perturbations
	Sarah Preheim JHU	Frequency and impact of transitions between microbial populations mediating biogeochemical cycling in a freshwater lake
	Nur A. Hasan CosmosID	Cloud based bioinformatics platform for cross-disciplinary microbiome research

535

536 **Supplementary Table 1.** Outline of oral presentations at the January 2018 M^3 Meeting

537

Break-out Session	Chairs	Topics
Viral and fungal pathogens: off the beaten path	Jacquelyn Meisel UMD CBCB and Daniel Nasko UMD CBCB	• Barriers to studying viruses and fungi and actionable goals to address • Detection of viral and fungal pathogens
Crowdsourcing biodefense: data sharing, data standards, data security, and data visualization	Todd Treangen UMD CBCB and Brian Ondov UMD, NHGRI, NIH	• Homomorphic encryption techniques for secure computations • Data visualization as incentive for data sharing • Data standards: raising the bar to ensure quality vs lowering the bar to bring data out of silos • Cloud based data sharing: practical or impractical?
Need for speed: navigating the trade-off between analysis accuracy and speed	Adam Bazinet NBACC and Nathan D. Olson UMD CBCB and NIST	• Evaluation of time-sensitive metagenomic analyses • Strategies for speeding up such analyses without compromising accuracy

538

539 **Supplementary Table 2.** Outline of interactive break-out sessions at the January 2018 M^3

540 Meeting

541

References

1. Drew TW, Mueller-Doblies UU. Dual use issues in research - A subject of increasing concern? Vaccine. 2017;35:5990–4.

2. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat. Rev. Genet. Nature Publishing Group; 2018;19:9–20.

3. Robinson ER, Walker TM, Pallen MJ. Genomics and outbreak investigation: from sequence to consequence. Genome Med. BioMed Central; 2013;5:36.

4. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for pathogen detection in public health. Genome Med. BioMed Central; 2013;5:81.

5. Lipkin WI. The changing face of pathogen discovery and surveillance. Nature Publishing Group. Nature Publishing Group; 2013;11:133–41.

6. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. Front Microbiol. Frontiers; 2017;8:1069.

7. main groups.io Group. Available from: https://m3.groups.io/g/main

8. Winter 2018 Mid-Atlantic Microbiome Meetup Biodefense and Pathogen Detection Agenda [Internet]. [cited 2018 May 4]. Available from: https://cpb-us-e1.wpmucdn.com/blog.umd.edu/dist/d/418/files/2017/10/WinterM3_agenda_final-27afpqx.pdf

9. CONSERVE: A Center of Excellence at the Nexus of Sustainable Water Reuse, Food, and Health, Year 1 Achievements (March 2016-February 2017) [Internet]. Available from: https://static1.squarespace.com/static/578101761b631b1a87aa0a3c/t/59f8f8e8e31d19ae528310e9/1509488877173/CONSERVE_annual_report.pdf

563 10. Singer E, Wagner M, Woyke T. Capturing the genetic makeup of the active microbiome in
564 situ. The ISME journal. Nature Publishing Group; 2017;11:1949–63.

565 11. Teytelman L, Stoliartchouk A, Kindler L, Hurwitz BL. Protocols.io: Virtual Communities for
566 Protocol Development and Discussion. Plos Biol. Public Library of Science; 2016;14:e1002538.

567 12. VERVE Net [Internet]. protocols.io. Available from: protocols.io/g/verve-net

568 13. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova
569 N, et al. Uncovering Earth's virome. Nature. Nature Research; 2016;536:425–30.

570 14. Yan J, Grantham M, Pantelic J, Bueno de Mesquita PJ, Albert B, Liu F, et al. Infectious virus
571 in exhaled breath of symptomatic seasonal influenza cases from a college community.
572 Proceedings of the National Academy of Sciences of the United States of America.
573 2018;115:1081–6.

574 15. Killingley B, Nguyen-Van-Tam J. Routes of influenza transmission. Influenza Other Respir
575 Viruses. Wiley/Blackwell (10.1111); 2013;7 Suppl 2:42–51.

576 16. Tellier R. Aerosol transmission of influenza A virus: a review of new studies. J R Soc
577 Interface. The Royal Society; 2009;6 Suppl 6:S783–90.

578 17. Bridges CB, Kuehnert MJ, Hall CB. Transmission of influenza: implications for control in
579 health care settings. Clin. Infect. Dis. 2003;37:1094–101.

580 18. Tellier R. Review of Aerosol Transmission of Influenza A Virus. Emerging Infect. Dis.
581 Centers for Disease Control and Prevention; 2006;12:1657–62.

582 19. Lemieux C, Brankston G, Gitterman L, Hirji Z, Gardam M. Questioning aerosol transmission
583 of influenza. Emerging Infect. Dis. 2007;13:173–4–authorreply174–5.

584 20. Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M. Transmission of influenza A in
585 human beings. The Lancet Infectious Diseases. Elsevier; 2007;7:257–65.

586 21. Conlan S, Park M, Deming C, Thomas PJ, Young AC, Coleman H, et al. Plasmid Dynamics
587 in KPC-Positive Klebsiella pneumoniae during Long-Term Patient Colonization. mBio.
588 2016;7:e00742–16.

589 22. Weingarten RA, Johnson RC, Conlan S, Ramsburg AM, Dekker JP, Lau AF, et al. Genomic
590 Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids Conferring
591 Carbapenem Resistance. Bonomo RA, editor. mBio. American Society for Microbiology;
592 2018;9:e02011–7.

593 23. Faust K, Lahti L, Gonze D, de Vos WM, Raes J. Metagenomics meets time series analysis:
594 unraveling microbial community dynamics. Curr. Opin. Microbiol. Elsevier Current Trends;
595 2015;25:56–66.

596 24. Gerber GK. The dynamic microbiome. FEBS Lett. Wiley-Blackwell; 2014;588:4131–9.

597 25. Bokulich N, Zhang Y, Dillon M, Rideout JR, Bolyen E, Li H, et al. q2-longitudinal: a QIIME 2
598 plugin for longitudinal and paired-sample analyses of microbiome data. bioRxiv. Cold Spring
599 Harbor Laboratory; 2017;:223974.

600 26. Cepeda V, Liu B, Almeida M, Hill CM, Koren S, Treangen TJ, et al. MetaCompass:
601 Reference-guided Assembly of Metagenomes. 2017.

602 27. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP - a flexible pipeline for genome-resolved
603 metagenomic data analysis. bioRxiv. 2018.

28. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology. BioMed Central; 2016;17:132.

29. Wagner J, Chelaru F, Kancherla J, Paulson JN, Zhang A, Felix V, et al. Metaviz: interactive statistical and visual analysis of metagenomic data. Nucleic acids research. 2018;514:59.

30. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature methods. Nature Publishing Group; 2017;14:1063–71.

31. Bremges A, McHardy AC. Critical Assessment of Metagenome Interpretation Enters the Second Round. mSystems. American Society for Microbiology Journals; 2018;3:537.

32. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat. Biotechnol. Nature Publishing Group; 2017;35:1077.

33. Altman RB, Prabhu S, Sidow A, Zook JM, Goldfeder R, Litwack D, et al. A research roadmap for next-generation sequencing informatics. Sci Transl Med. American Association for the Advancement of Science; 2016;8:335ps10–0.

34. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based species identification. bioRxiv. 2018.

35. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med. American Association for the Advancement of Science; 2012;4:148ra116–6.

626 36. Venkateswaran K, Singh NK, Checinska Sielaff A, Pope RK, Bergman NH, van Tongeren
627 SP, et al. Non-Toxin-Producing Bacillus cereus Strains Belonging to the B. anthracis Clade
628 Isolated from the International Space Station. Bik H, editor. mSystems. 2017;2:e00021–17.

629 37. van Tongeren SP, Roest HIJ, Degener JE, Harmsen HJM. Bacillus anthracis-Like Bacteria
630 and Other B. cereus Group Members in a Microbial Community Within the International Space
631 Station: A Challenge for Rapid and Easy Molecular Detection of Virulent B. anthracis. Schuch R,
632 editor. PLoS ONE. Public Library of Science; 2014;9:e98871.

633 38. Venkateswaran K, Checinska Sielaff A, Ratnayake S, Pope RK, Blank TE, Stepanov VG, et
634 al. Draft Genome Sequences from a Novel Clade of Bacillus cereus Sensu Lato Strains,
635 Isolated from the International Space Station. Genome Announc. American Society for
636 Microbiology Journals; 2017;5:e00680–17.

637 39. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial
638 Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. CELS. Elsevier;
639 2015;1:1–16.

640 40. Ackelsberg J, Rakeman J, Hughes S, Petersen J, Mead P, Schriefer M, et al. Lack of
641 Evidence for Plague or Anthrax on the New York City Subway. CELS. 2015;1:4–5.

642 41. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Modern
643 Methods for Delineating Metagenomic Complexity. CELS. 2015;1:6–7.

644 42. Bazinet AL, Ondov BD, Sommer DD, Ratnayake S. BLAST-based validation of
645 metagenomic sequence assignments. bioRxiv. Cold Spring Harbor Laboratory; 2017;:181636.

646 43. Langille MGI, Ravel J, Fricke WF. “Available upon request”: not good enough for
647 microbiome data! Microbiome. BioMed Central; 2018;6:8.

44. National Research Council. Science Needs for Microbial Forensics. Developing Initial International Research Priorities. Washington, D.C.: National Academies Press; 2014.

45. Bierer BE, Crosas M, Pierce HH. Data Authorship as an Incentive to Data Sharing. N. Engl. J. Med. 2017;376:1684–7.

46. Credit for Data Sharing [Internet]. 2018 [cited 2018 Aug 6]. Available from: https://www.aamc.org/initiatives/research/485818/datasharing.html

47. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. BioMed Central; 2008;9:386.

48. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7:335–6.

49. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology. American Society for Microbiology; 2009;75:7537–41.

50. ACSL.net [Internet]. [cited 2018 Aug 6]. Available from: http://ascl.net

51. Materials Resource Registry [Internet]. [cited 2018 Aug 6]. Available from: https://materials.registry.nist.gov

52. Hanisch RJ, Berriman GB, Lazio TJW, Emery Bunn S, Evans J, McGlynn TA, et al. The Virtual Astronomical Observatory: Re-engineering access to astronomical data. Astronomy and Computing. 2015;11:190–209.

670 53. Field D, Sterk P, Kottmann R, De Smet JW, Amaral-Zettler L, Cochrane G, et al. Genomic
671 standards consortium projects. Stand Genomic Sci. Michigan State University; 2014;9:599–601.

672 54. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic
673 Standards Consortium. Plos Biol. Public Library of Science; 2011;9:e1001088.

674 55. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum
675 information about a marker gene sequence (MIMARKS) and minimum information about any (x)
676 sequence (MIxS) specifications. Nat. Biotechnol. Nature Publishing Group; 2011;29:415–20.

677 56. Titus AJ, Flower A, Hagerty P, Gamble P, Lewis C, Stavish T, et al. SIG-DB: leveraging
678 homomorphic encryption to Securely Interrogate privately held Genomic DataBases. arXiv
679 Quantitative Methods. 2018.

680 57. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of
681 whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the
682 EUCAST Subcommittee. Clinical Microbiology and Infection. Elsevier; 2017;23:2–22.