Roundtable Discussion on Interrogation of Suspect Biological Samples: Capabilities and Problems Associated with Detecting Engineered Microorganisms and Deducing Function

Background - This paper reports on a September 15, 2016 Roundtable Discussion convened by B.Next, an IQT Lab. The purpose of the discussion was to explore whether and how a biological sample containing microorganisms could be examined using current techniques and procedures to answer two questions:

1) Is the sample likely to have been subject to genetic manipulation?
2) If the sample was engineered, is it possible to determine the intended and actual functionality of the genetic manipulation?

It was assumed that national security imperatives would impose some urgency on the need for information, so that the time required for different approaches is of concern. It was also recognized that the sample material available for examination might be limited and possibly irreplaceable. The source and type of biological sample at issue was not defined. Both clinical samples (presumably collected in the wake of an attack) and other types of collected samples, including complex, "metagenomic" samples (e.g. environmental effluents) were considered.

The Roundtable included twenty-six participants, including scientists from academia, seven US Government agencies and the Lawrence Livermore National Lab, representatives from private sector companies engaged in DNA sequencing and bioinformatics, and IQT professional staff. The group's expertise included bioinformatics, genetic engineering, computer science, microbiology, and biotechnology. The discussion took place over a single day, included invited presentations from three participants, and was held on a not-for-attribution basis.

Purpose – This Roundtable is part of the B.Next effort to design one of several IQT Lab Challenges. Challenge projects are proof-of-concept demonstrations that seek to probe the limits of capability or knowledge associated with a technology issue important to national security. This particular challenge – called GEMstone – is focused on describing the current US capacity to identify and characterize genetically engineered biological weapons. Based on this Roundtable discussion, IQT B.Next intends to pursue one or more GEMstone Challenge activities.

Summary of Discussion

Using current capabilities and available resources, it may be possible to detect genetic engineered microorganisms given substantial time (weeks). However, as genetic engineering becomes more widely used, particularly for private sector production using synthetic biology techniques, detecting clear "signals" of engineering will become more difficult.

Deducing the biological functions induced by genetic engineering remains a significant scientific challenge and would currently require substantial molecular experimentation (at least months of work) to provide specific predictions with confidence.

Participants agreed that no single “consensus protocol” for interrogating a biological sample is feasible due to the variability of sample type, resources, and expertise at any given time and place, and uncertainty about the assortment of tools and techniques that could be deployed. It would be possible to pre-determine the types of questions that would be posed and to identify the technological capabilities available to provide such answers.

For example, the combination of particular pathogens could point to intentional design and release because their combined effects on the human body are particularly lethal. So, one might ask “are there multiple species in this sample that are normally not found together and in combination present an unusually dangerous threat?” It will be critical to detect all pathogens in a sample, even if one is present in very small quantities compared to the others (i.e. <1%), to answer this question. This will mandate a particularly sensitive sample and data analysis approach – suitable for finding the “needle in the haystack.”

Discussion Topics

The Roundtable discussion covered policy topics (e.g. what questions do decision-makers need answered in what time frames?) as well as technical questions. This document offers a summary of the dialogue, reorganized into discrete subject areas.

Problems with Sample Interrogation

The discussants generally agreed that *metagenomic samples* – samples recovered directly from environmental samples and which contain multiple species of organisms – would be all but impossible to interrogate quickly, if at all. One discussant with experience in addressing this problem thought that a 30X sequencing depth was the minimum requirement needed for meaningful interrogation of a metagenomic sample. That is, the sample sequence would be generated at least 30 times to ensure sufficient accuracy. This level of sequencing already poses complex computation challenges in addition to the time and cost of sequencing samples with enormous backgrounds.

The problem of how to interrogate *“precious samples*” – limited amounts of biological material that could not be easily replaced - was raised. Determining the most efficient approach to interrogating such limited samples is clearly important, but there is no consensus agreement or standard on how to maximize information from limited samples. Precious samples (and probably all interrogations) would raise the question of whether attempts should be made to culture the specimen, which would increase material available for testing but would also delay analysis for days or more, and could also eliminate all non-culturable organisms contained in the original sample.

Two other major problem categories plague characterization of biological samples: the lack of accessible, comprehensive, and accurate genomic databases to which the sample can be compared; and the inadequate speed, scalability and accuracy of bioinformatics tools.

The comprehensiveness, quality and accessibility of databases containing sequences of relevant microorganisms are critical to sample analysis and currently inadequate.

Most analyses of genomic data are comparative, meaning what can be learned about a new, different and potentially important genomic sequence is based on what is already known about the genomic sequences of that microbial species or strain. Therefore, a repository of known genomic data available for comparative analysis is crucial.

Genomic data is held in a variety of public and private databases. Access to private data is limited by proprietary interests and, in some instances, classification barriers. The absence of widely accepted standards for DNA sequencing and data storage, the uneven quality of stored data, and variability in the inclusion of metadata (such as sequencing technique, assembly method, coverage, and error rates) are also problematic.

The three large, publically accessible databases include the US National Center for Bioinformatics (NCBI), the European Nucleotide Archive (ENA), and the Japanese equivalent of NCBI (DDBJ). While very valuable, all are incomplete, poorly curated due to insufficient funding, and difficult to navigate.

Private databases, held by individual researchers, institutions or private companies, are typically devoted to particular interests and access is limited. Genomic data held and generated by private companies are increasingly prolific and powerful. This trend towards privately held genomic data will strengthen as genome sequencing becomes central to research and businesses such as pharmaceutical companies, synthetic biology firms, etc. Two Chinese companies, BGI (formerly Beijing Genomics Inc.) and Novogene, are the world’s largest companies involved in DNA sequencing, and receive orders of large quantities of DNA from all over the world. They likely hold the largest compilations of genomics data.

It would be extremely valuable to establish one or more publically accessible genomic data clearinghouses, which could identify sources of useful genomic data. Classification issues and proprietary considerations that limit data access could be overcome by specific accessing agreements. For example, it might be possible to create privacy/proprietary data preserving inquiries by posting questions such as “do you have any genome sequences that look like this: “ACGT etc.?” However, an approach that also protects the sequence in question will have the highest likelihood of adoption with the government community.

There are many bioinformatics tools available to analyze genomic data, and more being invented daily, but they have limitations.

Discussion participants noted that bioinformatics algorithms are plentiful (“we have lots of tools”) but are generally unable to handle large datasets without “breaking.” In other words, the algorithms cannot produce results due to any combination of inadequate computational memory, inefficient data structures, and non-scalable algorithms whose required runtimes exponentially increase compared to input data size. Tools that were designed several years ago to handle a few tens of genomes need complete re-design to handle a few thousand (or more) genomes today. Mash[1] is a good example of a recent genomic and metagenomic analysis tool that demonstrates the kind of scalability required to handle the scale of today's and tomorrow's databases.

Many researchers develop their own algorithms, and more are published regularly, but the performance and limitations of these different algorithms are not integrated into a common framework or characterized in a way that enables rapid identification and deployment of which tools are most useful in specific situations. Also, new algorithms and bioinformatics tools are not catalogued, so scientists do not know who has what tools without doing a lot of time-consuming research.

Bioinformatics tools are typically developed for specific applications, making it difficult to grab and use these tools without alteration or customization. Moreover, bioinformatics tools, whether designed for general purposes or a customized application, are infrequently updated due to resource constraints, and may not remain viable in the context of computational advances.

***Detection of genetic engineering will not be straightforward and may not be possible with certainty*.**

Efforts to detect genetic engineering will likely start with attempts to align the suspect sample to known phylogenetic trees – a diagram showing inferred evolutionary relationships among species or entities based on genetic similarities and differences. The availability of robust genomic data for comparison will be crucial. The enormous variety of living organisms and the multitude of strains in some species in addition to the noted limitations of available databases will make this challenging.

One possible approach would be to seek evidence of unique or unusual genes not typically seen in the species of microorganism being examined. But one participant reported that the number of unique genes in already referenced genomes is considerable. For example, *B.anthracis* exhibits tens of unique genes per specific strain, but *F. tularensis* exhibits almost 200 unique genes per strain. Distinguishing an uncommon strain of *F.tularensis* from an engineered organism could prove extremely difficult.

Signals that organisms have been genetically engineered will get harder to detect. As more companies and researchers use synthetic biology techniques to create new products and develop new ways to engineer organisms, the variety of "signals" associated with engineering will evolve. One participant familiar with synthetic biology applications noted that increasingly sophisticated bioengineering, and the process of optimizing the desired bio-manufacturing process, results in organisms that appear more and more "natural". It

[1] Ondov *et al, Genome Biology* (2016) 17:132

was also noted that the unique processes used by different companies for DNA synthesis might leave unique signatures, which could make it possible to determine where the synthetic DNA came from.

Most participants agreed it would be difficult to detect CRISPR-edited genomes today. It is possible that evidence of gene editing might be left behind, but this is usually not the case, and researchers are already discovering new CRISPR-based gene editing enzymes, with more expected.

Determination of the intended function of an organism engineered as a bioweapon remains a science question.

There are no clear protocols for rapidly answering the question of the functionality of an engineered organism. Some participants believed "we could get a long way towards answering the question 'should we worry?" using current tools and knowledge. Most participants thought that functionality would be impossible to determine with confidence until more basic research is performed. Part of the difficulty is that important national security questions – e.g. what is this engineered organism designed to do? – have no counterpart or application in academic or private sector infectious disease research, and hence have not been pursued.

Some government agencies and laboratories have or are supporting efforts to address these issues, but funding is sporadic and limited to a handful of researchers. One National laboratory is investigating possible approaches to automated analysis of unknown genomic samples.

The challenge of determining engineered functionality would also confront some of the same limitations associated with identifying the sample in the first place: lack of access to needed genomic databases; inadequate or uncharacterized bioinformatics tools, etc. It is worth noting that our ability to computationally predict function of a novel protein or protein variant is inherently limited to our ability to compare against an annotated library of known proteins and their purported or experimentally validated function.

Another problem that would arise is the lack of integration of the government genomics community. Many of our government participants had not previously met and were unaware of each other's work. This is not a large community, but classification and diverse missions militate against collaboration and communication. In a national emergency crisis, it will be essential to make use of all the talent available to decision-makers, both within and outside of government, but today, there is not even a clear understanding of what expertise is available.

Opportunities to Improve Characterization of Unknown Biological Samples

Critical Questions. Participants noted that it would be useful to know, in advance of a crisis, exactly what questions decision-makers would want answered and in what time frames.

National Assets. It would be helpful to identify a national network that maps national capabilities and resources across the specialized fields, such as bioinformatics, within government, academia and the private sector. It may be prudent to ensure that a critical mass of scientists with the requisite skills have security clearances.

Data Sharing. Some discussants thought that the biggest obstacle to rapid and effective sample interrogation was scientists' reluctance to share data publically because of fear that they will lose professional credit for discovery and publication rights. If researchers are not incentivized to share data, it will be difficult to gather enough information to detect key patterns that might indicate engineering.

Raw-Read Analysis. Traditional bioinformatics analysis includes a pre-processing step ("assembly") of the raw data output ("raw reads") from genetic sequencing before the actual data analysis is performed. This pre-processing step often takes longer than the sequencing itself. Some suggested that using the raw reads for analysis instead of assembled genomes could speed results. This method is prone to some degree of error, however, so confidence in the results would be reduced. In situations where information is needed urgently and risk of error is acceptable, this type of analysis may offer the best solution. It was suggested that analysis with assembled genomes would have to follow to provide results with higher confidence.

Automated Analysis. Current analytic processes require a bioinformatics expert to manually set up the analysis, including selecting the parameters, thresholds, algorithms, and reference data. The continuous expansion of comparative genomic databases means that the analytics associated with querying these data become increasingly complex. Establishing automated protocols that could be immediately applied would significantly simplify and speed up the analysis process. Participants suggested some labs are already working on automated analysis pipelines to detect signatures of engineering but such work is in its infancy and poorly funded.

Machine Learning. Machine learning was cited as a new and potentially powerful approach to automated triaging of unknown biological samples. In view of the evolving libraries of genetic engineering signatures, a tool capable of learning and recognizing genomic patterns of engineering without requiring previous exposure to every possible combination would have tremendous value. This approach has yet to be proven in the genomics realm; its potential has been demonstrated in other applications, such as credit card fraud detection and filtering of email spam. Acquiring well-annotated training sets is the key obstacle for applying machine learning to problems of functional determination of unknown organisms or genes.

Crowdsourcing. The possibility of leveraging the greater scientific community to help detect evidence of engineered organisms and deduce their function was discussed. Such crowdsourcing has been successfully used in scientific research. "Foldit" was developed by University of Washington faculty to predict protein-folding patterns to enable better drug targeting. It has attracted over 50,000 "gamers" who regularly outperform computer algorithms.

To interrogate a suspect sample, the genomic sequence in question could be submitted to an established network of genomic experts, who would use their own tools and resources to analyze the data. The information exchange could occur over a web interface, and participation could be incentivized to maximize the expertise and resources available to interrogate the sample. Competitive incentives (e.g. prize money) have been used to speed results and increase confidence in other contexts.

The scientific community is generally eager to gain access to and analyze new pathogen sequence data. However, the release of new sequence data can be significantly delayed by the desire or need to publish results in the peer-reviewed literature. During an outbreak, this delay can postpone collaborative use of sequence data beyond the time when such collaborations could make real contributions to combatting the spread of disease.

Standardization. Creating standards for DNA sequence reporting and tool design would facilitate better integration of analysis methods and quicker ingestion of necessary data. The National Institute of Standards and Technology (NIST) is working closely with the FDA to establish standards for pathogen detection via next-generation sequencing.

Redteaming Events and Exercises. For any emergency situation (fire alarms, earthquakes, tornadoes, etc.), practice improves the speed and efficiency of the response. Nearly all participants suggested the genomics community and government agencies would greatly benefit from opportunities to practice rapid sample interrogation of unknown samples, such as redteaming events or full-scale exercises. Practicing on an ongoing basis (annually or biannually) would help maintain the institutional knowledge needed to act in a real emergency, even as experts rotate in and out of positions and technology evolves.

Conclusion –

The GEMstone Roundtable discussion strongly suggested that more could be done to prepare the government and the scientific community to rapidly and accurately interrogate biological samples suspected of being engineered bioweapons.

Although the problem of analyzing unknown samples has been considered and pursued by several groups in different federal agencies, the government as a whole is not ready to bring the sum of its expertise to bear on a potentially existential national security problem, nor is it prepared to rapidly engage top experts from universities and the private sector. Although the roundtable discussion was intentionally focused on technical and scientific challenges, it was apparent that the political challenges (e.g. no agency has "ownership" of the problem, much less than adequate funding, and lack of clearly defined mission responsibilities) are substantial and need to be addressed.

The scientific challenges behind this problem are significant. Ongoing progress in genomics and bioengineering will illuminate some of today's technical challenges, but these same forces will also propel more sophisticated biological engineering, and will make these technologies more widely accessible.

There are clear steps that could be taken to improve the Nation's readiness to interrogate unknown samples for evidence of bioengineering and to be better positioned to determine the function of engineered microorganisms. While it is to be hoped that the need to interrogate a possibly engineered bioweapon will never arise, the Roundtable discussion suggests that the problem of how to interrogate biological samples needs more attention. It would be prudent to pursue at least some of the preparatory actions, lest the country find itself unprepared at a time of great peril.