Workshop Report: Testing Sequence Screening

Date: December, 2022
Authors: Jacob Beal (Raytheon BBN), Sarah Carter (Science Policy Consulting), Adam Clore (Integrated DNA Technologies), Christina LaPosa (Raytheon BBN)

Executive Summary

Tools to identify controlled pathogen or toxin DNA sequences are crucial for biosecurity. Testing the efficacy of such tools is a complex challenge, however, and there are currently no shared materials or methods that can help support such testing.

To address this gap, we convened stakeholders for a Workshop on Testing Sequence Screening on November 15th, 2022, to discuss how shared test resources might be developed. Discussion focused specifically on nucleic acid and amino acid sequences controlled by the Australia Group Common Control Lists, US Commerce Control List, the US Select Agents and Toxins List, and the EU Dual Use List.

Key conclusions from this workshop are:

- There is general support in the stakeholder community for building shared test resources.
- Tests for controlled sequences have value, even if non-controlled sequences of concern are not covered.
- Multiple distinct classes of test have value, including:
    - tests at different levels of comprehensiveness directed at controlled sequences,
    - tests for accurate classification of non-threat sequences, including non-threat portions of threat organism genomes, and
    - tests directed at specific evasion techniques.
- Information hazards need to be managed and mitigated (particularly those related to evasion algorithms and to potential biorisk sequences in non-listed pathogens), but the test resources should be made as public as possible to enable input from a broad range of communities.
- High-quality annotations and metadata are valuable but expensive, so there is also value in using poorly annotated sequences to provide breadth (e.g. in variant coverage).
- Test resources will need ongoing maintenance, so their design needs to take sustainability and resource availability into account.

Motivation & Background

The ability to identify controlled pathogen or toxin DNA sequences is a critical requirement for effective biosecurity. In addition to being important for national security and to prevent accidental or malicious misuse of pathogens or toxins, identification of sequence content is also

required for legal reasons, so that an organization can ensure it is in compliance with export controls and other regulations. Many organizations are building tools for DNA sequence screening, whether open, commercial, or proprietary. At present, however, there is no consensus on how to test whether such a tool is effective at detecting controlled sequences.

Addressing this gap will require buy-in from diverse stakeholders with interests and expertise relevant to sequence screening and biosecurity more generally. If the stakeholder community is able to agree on a process to develop and maintain a shared set of materials and methods for testing, then such test resources are likely to be helpful in many contexts and for different communities. Note, however, there are many ways to think about what it means for a DNA sequence screening tool to be "good enough," and so test resources can only be the starting point for technical evaluation of a sequence screening tool, and not in and of itself a means of benchmarking or certification.

To begin to address this challenge, we organized a Workshop on Testing Sequence Screening to establish foundations for building shared community test resources. The workshop was held as a virtual meeting on November 15th, 2022, and was attended by an international group of stakeholders from academia, industry, government, and NGOs, including DNA synthesis providers, tool developers, regulators, pathogen experts, and biodefense / biosecurity experts, communicating under Chatham House Rule (i.e., information shared cannot be attributed to any specific participant). The workshop was organized with an initial session focused on the value of shared test resources and how they might be used, with framing remarks followed by a discussion. The second session focused on what sequences test resources should cover, with framing remarks followed by three break-out group discussions that were reported back to the whole group for broader discussion.

At the workshop, discussion focused on test resources designed to probe DNA sequence screening tools for false positives, false negatives, and robustness to evasion specifically for nucleic acid and amino acid sequences controlled by the Australia Group Common Control Lists, US Commerce Control List, the US Select Agents and Toxins List, and the EU Dual Use List. This report aims to serve as a roadmap for a community of practitioners to build and maintain shared test resources.

Summary of Workshop Discussions

The following descriptions aim to summarize key points that emerged from the discussion of workshop participants regarding each session topic.

Value Proposition for Shared Test Resources

The first session of the workshop asked the question, "Why do we need a common test set?" After some framing remarks from individual speakers, there was robust discussion about how test resources might be used and the value that common test resources would provide. Participants agreed that there are currently many sets of test sequences, but that they are being maintained by various different organizations and are not being widely shared. The

closest that currently exists to broadly shared test sets are two test sets used by the International Genome Synthesis Consortium (IGSC) for onboarding of new members. One of these test sets is a small set of controlled sequences and the other is a 1000-sequence test set containing mostly non-controlled sequences. Neither of these test sets aim to be comprehensive in any way, and both are significantly out of date.

Multiple different needs motivated participants to want shared test resources, and to want to contribute to their development. They believed that developing and maintaining shared test resources could be helpful for:

- Checking if a screening tool is working properly, either in isolation or as part of a larger business process (e.g., an ordering system).
- Sharing information about problematic "misfit" sequences (with either annotation issues or biological peculiarities) and about whitelists for "housekeeping genes" and other non-threat sequences found within threat taxa. This sort of information is currently maintained within individual organizations, but not shared between organizations.
- Sharing information about the control status of specific classes of sequences.
- Sharing information about known or proposed evasion methods and investigating potential defenses against them.
- Reducing the likelihood of harm from biological incidents.
- Reducing business risk through the ability to refer to shared best practices.
- Expanding the definition of sequences of concern beyond lists of controlled sequences (e.g., to functional categories such as immune subversion, adherence, invasion, and niche creation) in a way that is community-driven and defensible for individual companies.
- Supporting potential future certification or standards-setting processes.
- Better coping with emerging technologies and emerging shifts in what practitioners wish to synthesize.

Participants also identified a number of key issues that need to be overcome in developing shared resources, notably:

- Balancing breadth of input with concerns about information hazards. Openness and transparency would allow broader distribution of screening resources, particularly internationally, and would ensure that academic researchers (e.g. pathogen experts) can inform test resources. Information hazards can be minimized by, for example, listing only publicly available sequence records with well-established roles in pathogenicity. However, test resources that incorporate expert curation of potentially risky sequences, particularly those that are not well-documented, those not found in regulated pathogens, or those that can evade current detection systems, could possibly enable nefarious actors.
- Making sure that contributing to test resources does not create problematic legal or financial liability at a company for reasonable past decisions, e.g., at a company that may have, in the past, made reasonable, informed decisions on licensing requirements for shipped products that now conflict with the label given to an item in a test set.

- Ensuring that the "floor" for minimal performance of a sequence screening system does not also become a "ceiling" that users have no incentive to exceed.
- Avoiding overfitting, in which sequence screening tools "train to the test" or memorize its contents, and therefore do not perform as well in real-world contexts.
- Determining how to "grade" different hits and misses, considering that some organisms have much larger genomes than others (e.g., 1918 influenza vs. *Coccidioides immitis*), some sequences are much more dangerous threats than others (e.g., a toxin vs. a toxin-related regulatory sequence), some non-controlled materials are much more likely to be ordered than others (e.g., antibiotic resistance markers vs. poorly understood regulatory elements), and some organisms are much better studied than others.

Discussion of building test resources included both a "top down" and "bottom up" view of the problem. The "top down" view considers what should be tested in order to ensure a high degree of confidence and comprehensiveness. The complementary "bottom up" view considers the existing private test sets and how they might be shared and harmonized as a basis for improving on the current situation. Both views are important, and future development will need to balance between them based on resource availability.

A key idea that emerged from the discussion is that useful test resources are likely to be broader than simple sets of test sequences (though there is an important role for such test sets). Some test resources could be algorithmic in nature, such as programs that can generate test sets with specific characteristics (e.g. a set of sequences at a specified length with a specified percentage that match sequences of concern found in regulated pathogens) or more sophisticated algorithms that can generate novel instances of evasion methods or novel potential threats.

Scope of Testing

In the second session of the workshop, participants discussed what should be included in shared test resources. After some framing remarks, participants were invited to join one of three break-out group discussions: Coverage vs. cost (how comprehensive should it be?); Sequences to test robustness to evasion; and Granularity & classification (e.g., housekeeping genes, vaccine strains). Reports from these breakout sessions contributed to a final discussion with all workshop participants.

Comprehensiveness of Testing

One major question discussed regarding the development of test resources is how comprehensive the scope testing should attempt to be. Throughout the workshop, the focus was on test resources to support screening systems in identifying controlled sequences (though participants also noted that there would be significant value in test resources for sequences of concern that are not regulated). The number of sequences used to conduct a test needs to be limited due to the fact that screening is computationally expensive for some systems. Also, curation of sequences into test sets can be costly, and must be maintained over time to keep up with the changes in available information, potential threats, and regulatory requirements. It

was noted that in some cases larger test sets may actually be easier and cheaper to curate, since being less selective can mean requiring less expert hours of curation. Indeed, one potential extreme position is to simply include every known sequence without a known flaw (e.g., miscategorized or chimeric material) from a taxon of interest in the test set.

There was wide consensus amongst participants that different purposes of testing will require different levels of comprehensiveness. Key points included:

- Testing whether a screening tool has been deployed or installed correctly (e.g., as a component in a DNA synthesis ordering system) might be done with a small number of sequences, assuming that the tool has been more comprehensively tested during development.
- Certifying whether a screening tool is effective at discriminating controlled and non-controlled sequences within regulated taxa is likely to require a larger number of sequences that provide a representative sampling across the full breadth of these genomes, as well as "must-catch" evasions (e.g., variant strains, codon optimization), and significant categories of non-threats (e.g., common biotechnology tools, non-threat sequences closely related to threat sequences). Such a test set might be a generated subset of a larger set, in order to reduce the risk of overfitting.
- For development of screening tools and building consensus regarding unusual sequences and judgment calls on the boundary between controlled and non-controlled material, it is likely to be valuable to have much broader and more comprehensive coverage, including edge cases and problematic sequences.

Participants noted that there are significant open questions about how to define or measure comprehensiveness as well as how to balance the contents of any given set of test sequences in order to ensure adequate coverage. There was clear consensus that the interpretation of results was likely to be dependent on how a screening system is used: for example, some users of screening can tolerate significantly higher rates of false positives than others.

Other key questions not resolved in discussion included:

- How many sequences can screening systems afford to test?
- How much does the length of sequences tested affect computational load?
- How much do test sets need to grow or change over time?

Testing Robustness Against Evasion Methods

There was robust discussion of evasion methods that might be used to avoid detection by sequence screening systems and how to best test for these methods. Participants discussed different types of evasion, challenges to testing for them, and information hazards that may arise.

A number of "low hanging fruit" evasion methods are highly accessible via modern tools, to the point where "evasion" may often even be an unintentional side effect of ordinary bioengineering activities. Examples include codon optimization / codon shuffling, use of variant

strains, frame-shifting, and insertion of threat material into larger non-threat sequences. There was broad consensus that every screening tool should be capable of defending against such evasions, and thus tests addressing these evasions should be included in any testing regime.

A second class of evasion methods are known challenges that are more likely to appear only in the case of deliberate evasion attempts and which many tools have not yet been tested against. Examples of these evasion methods include use of close homologs from non-controlled taxa, splitting oligo orders between companies, reconstruction of threats via editing, and assembly of threats from short inserts via restriction digest. These sorts of evasion methods represent current challenges to screening, and would be valuable to include in testing in order to support evaluation and improvement of screening tools.

A third class of evasion methods require more sophisticated capabilities that are currently limited to a small number of actors, but which may become much more widespread in the future. Examples of these evasion methods include modification of the genetic code so that codons are interpreted differently, development of mirror-structure proteins, or *de novo* design of synthetic toxins. It is unclear how practical it is to test for such evasions in the near term, but there was general consensus that it would be valuable to investigate the potential risk posed by such threats and for those investigations to involve practitioners who are pushing the relevant scientific frontiers.

Evasion methods are typically combinatorial and generative, in the sense that there are many options for how to actually instantiate an evasion. For example, there are many possible alternative codon choices, many possible ways to split a sequence into oligos, and many possible alternative protein designs. As such, it is reasonable to have tests for evasion methods use randomized algorithmically generated sequences, rather than using a fixed test set of sequences. This can provide a value in avoiding the potential for overfitting, though repeatability of test results still needs to be assured.

An important caveat in evaluating the results of evasion tests, however, is that in many cases the functionality of the test sequence may not be known. For example, codon shuffling is known to disrupt expression in some cases, oligo decompositions may not assemble correctly, and predicted protein structure may not bear out in practice. Thus, a failure to detect a particular instance of evasion does not necessarily imply that it would be possible for an adversary to obtain a functional threat capability.

Participants also noted that there are unresolved regulatory questions around evasion methods. If a sequence constructed to evade detection provides access to the same capability as a regulated sequence to generate pathogenic or toxic effects, then one would expect that the evasion sequence should be subject to the same regulation. However, current list-based approaches to regulation do not necessarily extend regulatory requirements in this way, particularly with respect to *de novo* threats. It was also noted that it is possible that certain implementations of evasion methods might themselves fall under export regulation due to their ability to enable production of regulated materials. Development of shared test resources may

help make these issues more concrete and support community development of recommended practices beyond the technical scope of current regulations. Creation of a shared test set may also in and of itself create incentives for improving screening tools.

Information hazards are a critical question to address with respect to evasion methods. Many methods pose little threat, either because the methods are widespread and defenses already exist (e.g., codon shuffling) or because the methods are currently highly inaccessible (e.g., *de novo* toxin design). In some cases, however, developing test resources that can algorithmically generate evasion sequences may constitute a significant information hazard because screening tools may not defend against them and access to the test automation may greatly decrease the skill needed to utilize the evasion method. These risks will have to be carefully assessed before these testing tools are made available. It was also noted that risk management might be informed by examining how the software industry manages analogous information hazards regarding software security issues.

Length and Classification of Test Sequences

Non-threats and "clearable" materials in threat taxa are an important part of test resources because too many false positives in a screening system can impose significant costs and can lead to "decision fatigue" in which screening system users get accustomed to ignoring positive hits and therefore miss some true risk sequences. There was a clear consensus that one of the key benefits provided by shared test resources would be sharing information about sequences in controlled organisms that do not contribute their threat potential, and are thus not controlled.

Many participants believed that the commonly used notion of "housekeeping gene" is ill-defined and thus not particularly useful. There was a common understanding, however, that the majority of genetic material in listed bacteria and eukaryote threat taxa is non-controlled, by virtue of not endowing or enhancing pathogenicity (though this is not the case for viruses). The classification of test sequences can be used to capture and share this information.

There was general consensus that annotations and metadata will be a critical aspect of building useful test resources. It would be helpful to separate test sequences into multiple tiers of threat level with varying levels of evidence to support that classification (e.g., genes known to endow or enhance pathogenicity vs. genes merely unique to a regulated pathogen). Likewise, some genes can be safely categorized as being non-threatening, but the boundary between "safe" and "unsafe" genes is complex and difficult to determine. Many different methods for assessing the potential threat level of sequences were discussed, including:

- Judgment calls by subject matter experts
- Prediction by machine learning systems
- Identification of well-known non-threat functions
- Functional prediction from protein structure analysis
- Identification of host range
- Bioinformatic analysis of uniqueness

- Homology to structural features in other proteins

A major challenge in sequence classification is the cost of such determinations and other annotations, as well as the unbounded scope for other potentially useful information and methods for assessment. There was also a recognition that some categories of threat are easier to curate and automate than others, since some threats are simpler, less diverse, or more well-studied than others. For example, in small viruses, the function of every gene is often known, which is not the case for bacterial and eukaryotic pathogens, and human pathogens tend to be much better studied than agricultural pathogens.

Maintenance of testing resources was also identified as a major challenge in this area, since the constant expansion of sequence data means that test sets rapidly become out of date. Entangled with this is the question of who has the authority to alter the classification of a sequence, and what type of process would support ongoing changes to common test sets.

There was no consensus on how to select the size of test sequences. Genetic context is important for decision making in many cases, but may not be available to a screening tool. Shorter sequences are harder to classify, but there is no consensus on an appropriate threshold for the length of sequence that should be screened, and many are waiting to see how length thresholds change in the anticipated new US HHS guidance.

Participants also considered classification of threats that are regulated differently in different jurisdictions, or that are not currently regulated by any jurisdiction. Participants judged that if a class of sequence is regulated in any jurisdiction, then it would be valuable to include tests focused on those classes of sequence, as well as information about the relevant regulations. It was noted that no general curated list of regulatory agencies and regulations exists, outside of the small set tracked by the IGSC.

Likewise, there is value in providing tests focused on sequences of concern that are not regulated anywhere, such as genes that non-controlled pathogens use for immune subversion, adherence, invasion, or niche-creation. Such test resources, however, should be separable from evaluation of tool performance against the main body of controlled sequences.

Finally, it was observed that there would be value in developing a complementary set of tests for customer screening, but that such test resources would likely be outside of the scope of the effort being discussed in the workshop.

Conclusions

From the workshop discussions documented above, there are several key conclusions that appeared to have broad support from workshop attendees.

There was a clear consensus that shared test resources developed to test for identification of controlled sequences would be of significant value, and many participants are motivated to contribute to the development of these resources. There was also clear consensus that a suite

of tests rather than a single test set would best address the range of differing challenges and use cases that were identified (e.g. test sets with fewer or more sequences, tests directed at specific evasion techniques).

More comprehensive tests will allow increased confidence in screening tools, but never complete certainty due to the diversity of biological threats and potential evasion methods.

Tests need to include non-threat sequences as well as threat sequences in regulated taxa. There will be particular value in using these test resources to share information about non-threat portions of threat organism genomes, in order to reduce false positives.

High-quality annotations and metadata are extremely valuable, but due to the expense of producing them, there is also a place for poorly annotated data to provide breadth (e.g. in variant coverage).

Information hazards need to be managed and mitigated, especially for highly curated test sets and automation of evasion methods that are not yet handled well by screening tools. Test resources initially should be kept private, until information hazards can be assessed. However, it should be a goal to make these resources as widely accessible as possible in order to get broad input, particularly from subject matter experts, people who are pushing the boundaries of synthetic biology, and people thinking about potential evasion methods.

Test resources will need ongoing maintenance, so the design of test resources needs to include discussions of sustainability, resource availability, and processes for future decision making.

Workshop Participants

In addition to the organizers listed above, the following list of attendees have opted to have their presence as workshop participants be acknowledged:

- Craig Bartling (Battelle)
- Kathryn Brink (Stanford University)
- Audrey Cerles (Gryphon Scientific)
- Ronald Coleman (Raytheon BBN)
- Michael Daniels (Evonetix Ltd)
- James Diggans (Twist Bioscience)
- Margaret Dunbar (Codex DNA)
- Nathan Dwarshuis (National Institute of Standards and Technology)
- Steve Evans (BioMADE)
- Kevin Flyangolts (Aclid, Inc.)
- Robert Friedman (J. Craig Venter Institute)
- Bryan Gemler (Battelle)
- Thomas Hofmeister (Thermo Fisher Scientific)
- Charles Hong (DTRA)
- Chris Isaac (NTI)
- Wesley Johnson (Bureau of Industry and Security, US Department of Commerce)

- Becky Mackelprang (Engineering Biology Research Consortium)
- Brittany Magalis (Nuclear Threat Initiative | Bio)
- Jeff Manthey (Integrated DNA Technologies)
- Tom Mitchell (Raytheon BBN)
- Steven Murphy (Raytheon BBN)
- Sophie Peresson (DNA Script)
- Lisa Simirenko (Joint Genome Institute)
- William So (US Government)
- Peter Vegh (University of Edinburgh)
- Beth Vitalis (Unaffiliated)
- Nicole E. Wheeler (University of Birmingham)
- Justin Zook (National Institute of Standards and Technology)