The Case for Limiting 'Sequences of Concern' to Those with Demonstrated Pathogenic Function

Gene Godbold[1*] (0000-0002-5702-4690), Krista Ternus[2] (0000-0003-1138-5308), Kevin Flyangolts[3] (0009-0000-5232-9615), Nicole Wheeler[4] (0000-0003-4599-9164), Michael Parker[5] (0000-0002-4359-8677), Jacob Beal[6] (0000-0002-1663-5102), Peter A Carr[6] (0000-0002-4078-8846), Kemper Talley[6] (0009-0003-0025-158X), Caitlin Jagla[6] (0000-0002-6139-8150), Bryan Gemler[7] (0000-0002-9635-4545), Craig Bartling[7] (0000-0002-3040-7149), Rebecca Mackelprang[8] (0000-0002-2851-3823), India Hook-Barnard[8] (0000-0002-7103-2152), James Diggans[9] (0000-0003-3871-2380), Samuel P. Forry[10] (0009-0000-8200-0327), Sheng Lin-GIbson[10] (0000-0001-5092-1519), Tyler Laird[10] (0000-0003-2317-2895), Todd Treangen[11] (0000-0002-3760-564X), Tessa Alexanian[12] (0000-0003-1108-7124), Jens Berlips[13] (0009-0007-3429-9265), Gregory Koblentz[14] (0000-0002-0793-1209), Kevin Esvelt[15] (0000-0001-8797-3945), Joshua Gil[1] (0009-0009-4866-001X)

Affiliations:
[1] Signature Science, LLC; 1670 Discovery Drive, Charlottesville VA 22911 USA
[2] Signature Science, LLC; 8329 MoPac Expy, Austin, TX 78759 USA
[3] Aclid, Inc; 442 5th Ave 2300, New York, NY 10018 USA
[4] Institute of Microbiology and Infection, University of Birmingham, UK
[5] Georgetown University; 3700 O St. NW, Washington, DC 20057 USA
[6] RTX BBN Technologies; 10 Moulton Street, Cambridge MA 02138 USA
[7] Battelle Memorial Institute; 505 King Avenue, Columbus OH 43201 USA
[8] Engineering Biology Research Consortium; 1900 Powell St Ste 1200, Emeryville, CA 94608 USA
[9] Twist Bioscience, 681 Gateway Blvd, South San Francisco, CA 94080 USA
[10] NIST, 100 Bureau Dr. Gaithersburg, MD 20899 USA
[11] Department of Computer Science, Rice University, Houston, TX 77251 USA
[12] International Biosecurity and Biosafety Initiative for Science (IBBIS), Geneva, Switzerland
[13] SecureDNA Foundation; Aeschenplatz 6, Basel, Switzerland 4010
[14] Schar School of Policy and Government, George Mason University, Arlington, VA, 22201 USA
[15] Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA
*Corresponding author: ggodbold@signaturescience.com

Author Contributions:

Conflict of Interest Declaration:

Ethical Compliance: The preparation of this article involved no human or animal research.

Keywords: synthetic DNA screening, sequence of concern,

Word Count: 2211

Abstract

Screening synthetic nucleic acid orders for sequences of concern is a necessary part of a healthy biosecurity regime, but it exacts costs for nucleic acid providers. Taxonomy is and will remain a critical part of the decision-making process for screening, especially for viral sequences. But, moving forward, the function of a sequence will also be determinative of its level of concern, or lack thereof. Stakeholders continue to debate which functions are "of concern." But however these are ultimately adjudicated, non-viral sequences with unknown or hypothetical functions which, by definition, can bear no resemblance to sequences with concerning functions, must be considered innocent of harmful effects. To qualify as a non-viral sequence of concern, the sequence to which it is a best match must be demonstrated in the published literature to have a function of concern.

Introduction

Many providers of synthetic nucleic acids have been screening their orders for sequences from regulated organisms for over a decade. The basis of this screening has primarily been on the assessed taxonomic origin of the sequence with those from regulated biological agents being subject to further review. The taxonomic lists were formulated on the demonstrated or predicted ability of the listed pathogen to be employed as a biological weapon. Such screening has never been required and relies upon voluntary, good faith efforts, mainly orchestrated through the International Gene Synthesis Consortium (IGSC)(1). In current practice, sequence screening is combined with establishing customer identity and legitimacy to reduce the likelihood of misuse of these sequences.

The development of the ability to manipulate DNA sequences via restriction enzymes led to the 1975 Asilomar Conference on Recombinant DNA. The gathered researchers were concerned about hazards that involved combining genes from different organisms and propagating them in non-native biological systems. Many were worried about cancer viruses. Some pondered the dangers of 'improved' infectious agents turned into biological weapons (2,3).

Many of the biohazards considered during that period have proven less dangerous than originally thought. The principal exception is the threat arising from infectious agents of humans and the agricultural species on which society relies. Earlier conceptions of threat biology focused entirely on the taxonomic identity of microbes which could be identified by laboratory techniques that focused on biochemical proxies of the agent. Many interested parties have recognized that screening nucleic acids for biosecurity must move beyond taxonomic lists (4). Some have focused on the *function* of a sequence from a human pathogen as a basis for regulation and screening guidance, specifically on those sequences that the pathogen employs to manipulate and subvert host biology to dysregulate host homeostasis (5,6). These sequences are those that have historically been investigated by researchers in the field of microbial pathogenesis (7,8).

The first microbial pathogenesis investigations that involved transfer and removal of genes were those of Williams Smith and Margaret Linggood who published on enteropathogenic

Escherichia coli (EPEC) in pigs in the early 1970s. They showed that a bacterium that could not cause disease could be reliably converted to one that does by the addition of particular (Ent) plasmids (9). Similarly, loss of a plasmid from a diarrheagenic strain resulted in the modified bacteria losing the capacity to cause disease (10). These studies demonstrated that a particular enterotoxin sequence could endow a microbe with pathogenic capacity.

A growing armamentarium of molecular biological tools, developed since the 1970s, have powered tens of thousands of explorations seeking to unveil the function of thousands of sequences used by pathogenic microbes—bacterial, viral, fungal, and protozoal—in exploiting their hosts. It is on these sequences that regulations and screening should be particularly focused as these have the greatest potential to cause harm if engineered into new or existing pathogens. These are the sequences that, when expressed effectively, damage hosts, subvert and counter host innate immunity, allow microbial dissemination, and manipulate host cellular processes. These are certainly "sequences of concern" (SoCs) (5,6,11). There are probably other sequences that could be considered SoCs, perhaps even from host taxa (12), but the overwhelming majority will be from pathogenic species that have adapted to exploit host biology, suppressing innate immunity and evading adaptive immunity to increase transmission and, perhaps, virulence.

Sequences with Unknown or Hypothetical Functions from Non-viral Pathogens Are <u>Not</u> Concerning When Screening Synthetic Nucleic Acids

The Framework for Nucleic Acid Synthesis Screening issued in April 2024 by the White House Office of Science and Technology Policy (OSTP) indicates that, as of October 13, 2026, a sequence of concern will include those "known to contribute to pathogenicity or toxicity, even when not derived from or encoding regulated biological agents". Moreover, "known to contribute" means that the "direct and harmful impact on a host has been verified based on published experimental data; and, where experimental data do not exist, based on homology to a sequence encoding a verified function" (13). As of October 2026, a sequence without the ability to interact deleteriously with a host cell or tissue is not of concern. This function with reference to a host species must be present in the literature for that sequence to be "of concern". The updated Framework wisely anchors the sequence function to published investigations in which those functions are revealed by empirical research as well as sequences that strongly resemble the investigated (original) sequence. If there were anything we could add to the updated Framework, then we would specify that the research should require positive evidence of pathogenicity for a given sequence for a given host and not merely show diminished pathogenicity as a result of deletion or alteration of the sequence.

Documenting the host-exploiting function of sequences from many different microbes (5,6) has allowed us to recognize an important feature of non-viral pathogens: the sets of sequences that enable them to interact with host biology comprise a small fraction of the pathogen genome. We previously enumerated the SoCs in SARS-CoV-2 and compared them to those of *Bacillus anthracis* Of the 27 proteins expressed by SARS-CoV-2, all but three directly exploit the host;

eighteen are involved in subversion of innate immunity (11). Viral genomes are the most compact of pathogen genomes with nearly all the sequences devoted to host exploitation. For this reason, viral taxonomy is a much better proxy for concerning functions than it is for bacterial and eukaryotic pathogens. In contrast, there are fewer than two dozen SoCs with documented pathogenic functions out of ~5700 protein coding genes in the genome of *Bacillus anthracis*: just 0.35% of the genome encodes SoCs.

The virulence of *Bacillus anthracis* has been thoroughly investigated so we had some confidence its genome holds no pathogenic surprises. One can only annotate what has been published. Annotation always lags behind published work and many pathogens have not been examined for anything to be published on their host-exploiting sequences. The case of *Legionella pneumophila* is illustrative. Ten percent of its genome, 300 genes, are involved in host manipulation (14). This is believed to be the highest percentage of any bacterial pathogen devoted to pathogenesis, but functions for less than half of them have been elucidated by researchers. **Table 1** compares annotated SoC content with genome size for some bacterial pathogens of humans, a fungal pathogen (*C. albicans*) and a few near neighbors, most of which are opportunistic pathogens.

Microbial Species	Genomes	Max SoC	Min SoC	Mean SoC	SoC %	Mean Gene Count
Bacillus anthracis	127	20	15	19.1	0.34	5689
Bacillus cereus	134	19	13	16.8	0.29	5844
Bacillus cytotoxicus	17	4	4	4	0.09	4471
Bacillus mycoides	53	17	12	15.2	0.25	6035
Bacillus pseudomycoides	4	4	3	3.5	0.06	5927
Bacillus thuringiensis	95	20	0	17.1	0.27	6316
Burkholderia mallei	33	6	2	5.8	0.11	5209
Burkholderia pseudomallei	135	7	6	6.7	0.10	6449
Burkholderia thailandensis	24	6	4	5.5	0.09	5987
Burkholderia vietnamiensis	10	0	0	0	0	5968
Candida albicans	1	5	5	5	0.08	6263
Candida tropicalis	1	1	1	1	0.02	6441
Clostridioides difficile	159	4	0	2.7	0.07	3835
Clostridium butyricum	17	1	0	0.1	0.002	4247
Clostridium perfringens	123	4	1	1.9	0.06	3127
Francisella novicida	1	19	19	19	1.03	1841
Francisella philomiragia	6	4	4	4	0.20	1983
Francisella tularensis	61	19	7	16	0.82	1948
Haemophilus influenzae	108	5	0	1.8	0.01	1879
Legionella pneumophila	145	142	82	122	3.96	3080
Salmonella enterica	1741	53	19	43.8	0.91	4811
Staphylococcus aureus	1502	45	19	32.7	1.15	2840
Vibrio cholerae	140	13	3	9.7	0.26	3779
Yersinia pestis	136	21	6	17	0.39	4342
Yersinia pseudotuberculosis	34	19	8	13.3	0.31	4329

Table 1: **Detected SoCs in select, completed Refseq genomes**. This shows an analysis of 4817 genomes from 25 microbial species collected from Refseq (https://www.ncbi.nlm.nih.gov/refseq/). Manually curated SoCs were searched against each genome using tblastn. Unique SoC hits above 90% sequence identity and a bit score greater than 80% of the original were cataloged for every genome. Shown are the total number of genomes per microbial species analyzed, the

maximum and minimum number of SoCs detected in strains within that species, the mean SoCs detected, the % of mean SoCs in the average genome and the mean protein-coding gene count for each species. Species highlighted in red are pathogenic for immune competent humans while species highlighted in gray are opportunistic pathogens. Certain strains of *Clostridium butyricum* can encode botulinum neurotoxin E but the species is not generally considered a pathogen.

Sequences with Unknown and Hypothetical Functions Cannot Be Considered Sequences of Concern

Proteins which only have homeostatic and regulatory functions within a non-viral microbe (i.e. 'benign' sequences) are not sequences of concern, even when encoded by a regulated pathogen. But what about sequences that only have non-threatening hypothetical functions or those whose functions are completely unknown? Proteins of indeterminate function comprise a significant portion of all extant genes discovered.

Approximately 40-60% of predicted genes have unknown functions, and that number has been increasing with accumulating sequence data (15–17). Concerted efforts to understand the function of unknown human proteins has reduced the percentage from 43% to 23% over the past 10 years. But there has not been a similar decrease for unknown proteins of non-model organisms (18). Unknown proteins that share sequence homology are together assigned a domain of unknown function (DUF). In the Pfam database (v. 35.0), nearly 4,800 of the 19,632 entries (24%) are grouped into a DUF (19). DUFs are challenging to identify and are not usually a primary focus for researchers.

However, DUFS are obviously biologically relevant. In 43 bacterial species in the Database of Essential Genes (20), there were 404 proteins that contained at least one of 297 different DUFs as of November 2024. However, these sequences cannot be considered sequences of concern because they are not known to contribute to pathogenicity or toxicity.

For providers of nucleic acid sequences, the two criteria of a sequence of concern are (i) that it can endow a microbe with pathogenic or toxic capacity and (ii) that this pathogenic or toxic capacity should be supported by experimental evidence. These requirements would keep the scope of the screening task tractable, though not simple. Requiring providers to refuse, and potentially report, orders for sequences with hypothetical and unknown functions, even if they are from regulated microbes, not only increases the workload but puts them in an untenable situation with regard to their customers. There is no reason to believe these sequences, beyond mere provenance, are dangerous. The provider cannot know something that isn't known and has no basis for neglecting to supply the sequence. For unknown sequences from unlisted microbes, not even the provenance can be used as a justification to 'have a conversation' with the customer.

Conclusion

While sequences of unknown and hypothetical function, even from regulated pathogens, do not require reporting in sequence screening efforts, this is not the case for sequences that enable toxicity and pathogenesis in human pathogens whether regulated or not. It would be easiest for synthetic nucleic acid sequence providers if governments could devise a standard list of SoCs following consultations with experts. How this list would be selected, maintained, and used is something that will need to be resolved. The first question for such a group involves selecting which host taxa require protection. Humans are the primary concern, but animals and plants dominating a country's agriculture could be considered. Establishing the requisite hosts needing protection allows the selection of pathogen species from which SoCs would then be drawn.

The availability of such a list of SoCs and the type of information it provides is also something to be decided. Should it be an open list of sequence names? A list of names with accession numbers? The names, accession numbers, and a tabular list of problematic functions and/or controlled vocabulary terms? Should citations from primary and secondary literature be required to justify the selection of each sequence?

Who should have access to such a list? Should it be public or available only to institutional biosafety committees and the businesses that need to screen for SoCs? Should different groups have access to lists of differing comprehensiveness? The utility of the tool for screening sequences needs to be balanced with the information hazards presented by an accessible compilation of sequences that enable pathogenesis. Those making the decisions will need to skillfully discriminate among the goods of public safety, open research, and international security.

Citations

1. Wheeler NE, Bartling C, Carter SR, Clore A, Diggans J, Flyangolts K, et al. Progress and Prospects for a Nucleic Acid Screening Test Set. Appl Biosaf. 2024 Sep;29(3):133–41.
2. Berg P, Baltimore D, Brenner S, Roblin RO, Singer MF. Summary statement of the Asilomar conference on recombinant DNA molecules. Proc Natl Acad Sci U S A. 1975 Jun;72(6):1981–4.
3. National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Board on Life Sciences, Board on Chemical Sciences and Technology, Committee on Strategies for Identifying and Addressing Potential Biodefense Vulnerabilities Posed by Synthetic Biology. Biodefense in the Age of Synthetic Biology [Internet]. Washington (DC): National Academies Press (US); 2018 [cited 2024 Dec 4]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK535877/
4. Millett P, Alexanian T, Brink KR, Carter SR, Diggans J, Palmer MJ, et al. Beyond Biosecurity by Taxonomic Lists: Lessons, Challenges, and Opportunities. Health Secur. 2023 Oct 19;
5. Gemler BT, Mukherjee C, Howland CA, Huk D, Shank Z, Harbo LJ, et al. Function-based classification of hazardous biological sequences: Demonstration of a new paradigm for biohazard assessments. Front Bioeng Biotechnol. 2022;10:979497.
6. Godbold GD, Kappell AD, LeSassier DS, Treangen TJ, Ternus KL. Categorizing Sequences of Concern by Function To Better Assess Mechanisms of Microbial Pathogenesis. Infect Immun. 2022 May 19;90(5):e0033421.
7. Falkow S. Molecular Koch's postulates applied to microbial pathogenicity. Rev Infect Dis. 1988 Aug;10 Suppl 2:S274-276.

8. Falkow S. Molecular Koch's postulates applied to bacterial pathogenicity--a personal recollection 15 years later. Nat Rev Microbiol. 2004 Jan;2(1):67–72.
9. Smith HW, Linggood MA. The transmissible nature of enterotoxin production in a human enteropathogenic strain of Escherichia coli. J Med Microbiol. 1971 Aug;4(3):301–5.
10. Smith HW, Linggood MA. Observations on the pathogenic properties of the K88, Hly and Ent plasmids of Escherichia coli with particular reference to porcine diarrhoea. J Med Microbiol. 1971 Nov;4(4):467–85.
11. Godbold GD, Hewitt FC, Kappell AD, Scholz MB, Agar SL, Treangen TJ, et al. Improved understanding of biorisk for research involving microbial modification using annotated sequences of concern. Front Bioeng Biotechnol. 2023;11:1124100.
12. Jackson RJ, Ramsay AJ, Christensen CD, Beaton S, Hall DF, Ramshaw IA. Expression of mouse interleukin-4 by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mousepox. J Virol. 2001 Feb;75(3):1205–10.
13. Office of Science and Technology Policy (OSTP). Framework for Nucleic Acid Synthesis Screening [Internet]. 2024 Apr [cited 2024 Jun 18] p. 13. Available from: https://www.whitehouse.gov/wp-content/uploads/2024/04/Nucleic-Acid_Synthesis_Screening_Framework.pdf
14. Omotade TO, Roy CR. Manipulation of Host Cell Organelles by Intracellular Pathogens. Microbiol Spectr. 2019;7(2).
15. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019 Apr;568(7753):499–504.
16. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. Genome Biol Evol. 2018 Mar 1;10(3):707–15.
17. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018 Jan 25;9(1):373.
18. Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, et al. Functional unknomics: Systematic screening of conserved genes of unknown function. PLoS Biol. 2023 Aug;21(8):e3002222.
19. Lv P, Wan J, Zhang C, Hina A, Al Amin GM, Begum N, et al. Unraveling the Diverse Roles of Neglected Genes Containing Domains of Unknown Function (DUFs): Progress and Perspective. Int J Mol Sci. 2023 Feb 20;24(4):4187.
20. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res. 2021 Jan 8;49(D1):D677–86.