Godbold *et al.*
Journal of Biomedical Semantics (2025) 16:4
https://doi.org/10.1186/s13326-025-00323-8

Journal of Biomedical Semantics

RESEARCH

Open Access

New and revised gene ontology biological process terms describe multiorganism interactions critical for understanding microbial pathogenesis and sequences of concern

Gene Godbold[1*], Jody Proescher[2] and Pascale Gaudet[3*]

Abstract

Background There is a new framework from the United States government for screening synthetic nucleic acids. Beginning in October of 2026, it calls for the screening of sequences 50 nucleotides or greater in length that are known to contribute to pathogenicity or toxicity for humans, regardless of the taxa from which it originates. Distinguishing sequences that encode pathogenic and toxic functions from those that lack them is not simple.

Objectives Our project scope was to discern, describe, and catalog sequences involved in microbial pathogenesis from the scientific literature. We recognize a need for better terminology to designate pathogenic functions that are relevant across the entire range of existing parasites.

Methods We canvassed publications investigating microbial pathogens of humans, other animals, and some plants to collect thousands of sequences that enable the exploitation of hosts. We compared sequences to each other, grouping them according to what host biological processes they subvert and the consequence(s) for the host. We developed terms to capture many of the varied pathogenic functions for sequences employed by parasitic microbes for host exploitation and applied these terms in a systematic manner to our dataset of sequences.

Results/Conclusions The enhanced and expanded terms enable a quick and pertinent evaluation of a sequence's ability to endow a microbe with pathogenic function when they are appropriately applied to relevant sequences. This will allow providers of synthetic nucleic acids to rapidly assess sequences ordered by their customers for pathogenic capacity. This will help fulfill the new US government guidance.

Keywords Controlled vocabularies, Dual use research of concern, Infectious diseases, Microbial pathogenesis, Ontologies, Synthetic nucleotide screening

*Correspondence:
Gene Godbold
ggodbold@signaturescience.com
Pascale Gaudet
pascale.gaudet@sib.swiss
Full list of author information is available at the end of the article

Background

In April 2024, a new framework for nucleic acid synthesis screening was issued by the White House Office of Science and Technology Policy (OSTP). The purpose was "…to encourage providers of synthetic nucleic acid sequences to implement comprehensive, scalable, and verifiable synthetic nucleic acid procurement screening mechanisms…[to] minimize the risk of misuse" [1]. This framework formalized, as a condition of receiving life sciences research funding from the US government, the October 2023 guidance from the Department of Health and Human Services (DHHS) to Providers and Users of Synthetic Nucleic Acids [2]. That framework calls for providers of synthetic nucleic acids to screen ordered sequences 200 nucleotides or longer to identify sequences of concern (SoCs). Until October 2026, SoCs are defined as sequences that are best matches to a sequence of the Biological Select Agents and Toxins List or, for international orders, the Commerce Control List except when the sequence is also found in an unregulated organism or toxin. In October 2026 the sequence length to be screened will decrease to 50 nucleotides and SoCs will be defined as sequences known to contribute to pathogenicity or toxicity, even when not derived from regulated biological agents [1].

The phrase "known to contribute" is described, in both the OSTP and DHHS policies, as requiring published experimental data and, where such data is lacking, based on similarity ("best match") to a sequence encoding a verified pathogenic or toxic function. We have been involved in efforts to collect and describe the roles of these sequences in pathogenesis and toxicity for programs funded by the United States Government, including the Functional Genomic and Computational Assessment of Threats (Fun GCAT) program of the Intelligence Advanced Research Project Activity (IARPA) [3]. From 2017 to 2022, the computational portion of the Fun GCAT program funded tool development to answer three questions about any given sequence: (1) what is the original taxon? (2) what are its biological functions? and (3) how dangerous is it? We observed that the danger inherent in a sequence from a pathogen was dependent on what anti-host functions it possessed. When considering the intelligent design of a microbe engineered for human woe, if a sequence had a concerning function in one organism, then it seemed reasonable to assume that function could be maintained when transferred to another organism. Through extensive review of the literature, we came to realize that 'dangerous' sequences were found primarily among virulence factors from pathogenic microbes and secondarily from venom-producing taxa.

Pathogenic microbes—viruses, bacteria, fungi, and protozoa—cause an array of human diseases. Pathogens are distinguished from their nonpathogenic relatives, including nonpathogenic strains of the same species, by specific molecules (carbohydrates, lipids, proteins and combinations thereof as well as small RNAs) that endow them with the capacity to exploit particular hosts [4, 5]. These sequences, often called virulence factors, play essential roles in infectious disease [6].

As we attempted to categorize 'dangerous' sequence functions, we found that, for nonviral and nontoxic sequences, there was no existing, standardized terminology to distinguish those that were concerning from those that were innocuous [7]. To remedy this, we developed two controlled vocabularies: Functions of Sequences of Concern (FunSoCs) [7, 8] and the Pathogenesis Gene Ontology (PathGO) [9], then used these to label sequences we had annotated from the published literature in microbial pathogenesis. Other groups were also working on this in parallel [10].

Later, we realized a more universally applicable set of terms was necessary to annotate sequences in public databases. While FunSoCs were not meant to be a comprehensive set of descriptions, PathGO was established to provide more granularity. Though it proved unsuitable for annotating public datasets, PathGO did inform our revision and expansion of the gene ontology (GO) terms applicable to microbial pathogenesis. This article briefly discusses FunSoCs and PathGO then analyzes our work adapting GO biological process terms to make them pertinent to pathogenic functions of microbes.

FunSoCs

FunSoCs were a quick solution for our sequence-gathering effort that began in 2018. As we have described elsewhere, at that time there was no available controlled vocabulary for denoting nonviral pathogenic sequence activity [7, 11]. We binned sequences according to what host biological process was affected: transcription, translation, cell cycle, cytoskeleton dynamics, the endomembrane system, autophagy, regulated cell death, small GTPases serving as molecular switches, and ubiquitination.

We also attempted to capture the pathogenic consequences of the sequence activity. We categorized sequences involved in adherence to host molecules, and invasion of the host, and if they enabled active dissemination of the microbe through host barriers. Intracellular pathogens often hijack host cellular components to develop a protected replicative compartment, which is also a FunSoC category.

Sequences from pathogens that damage the host were binned according to whether they (1) disabled a host organ (2) lysed or otherwise killed the host cell (3) permeabilized tissue structures or (4) caused inflammation.

The cause-and-effect for sequences associated with inflammation can be particularly difficult to discern. The natural consequence of the host detecting a microbial component or microbial activity disrupting host homeostasis (translation blockage, cytoskeleton disruption, organelle stress, etc.) is activation of pathways that result in inflammatory damage to the host [12, 13]. This is a host-directed, evolved activity. But a few pathogen sequences enzymatically activate host signaling pathways to force an inflammatory response [14–16]. We had the most, and the most varied, FunSoCs for sequences that affected host innate immunity. If the sequence altered a microbial molecule so it was less detectable by a host sensor, then it was categorized as passive immune evasion. If the microbial sequence inactivated or disrupted a host immune component or immune effector, then it was categorized as immune subversion.

At the time of their generation, we recognized that the ~ 30 FunSoC terms we developed did not provide a sufficiently granular description of sequences of concern. They were nevertheless useful for denoting consequences of the action of microbial virulence factors during pathogenesis that it was otherwise hard to capture. Many are helpful as grouping terms [7, 8], though some, such as "subverting host innate immune signaling" were intolerably broad as they encompassed dozens of discrete cellular signaling pathways (see Table 5 below).

Ontologies for the life sciences—representing microbial pathogenesis

Ontologies are structured vocabularies that define concepts – and relationships between concepts – in a particular domain. Their structure allows computers to reason over them, and the biomedical informatics community has long recognized the utility of ontologies in aggregation and analysis of complex data [17]. Microbial pathogenic processes differ from homeostatic processes occurring in a single organism as the activity of toxins and virulence factors involve at least two sequences from different organisms interacting in the space of the target organism, posing a challenge for ontological representation. While GO represents inter-species interactions [18, 19], the terms are broad and have lacked specificity for pathogenicity over mutualistic interactions [20]. They have also failed to differentiate intra-organism (homeostatic) versus inter-organism effects (mutualistic or pathogenic). PathGO was an initial attempt to incorporate pathogenic interactions between microbes and hosts.

PathGO

PathGO was developed as an application ontology describing mechanisms of pathogenesis to support the straightforward, unambiguous annotation of viral, eukaryotic, and bacterial genes. This application ontology fills a previously recognized gap for a focused ontology of to improve sequence annotation related to mechanisms of pathogenesis [20]. PathGO has been maintained in a public source code repository on GitHub (https://github.com/jhuapl-bio/pathogenesis-gene-ontology) [9]. PathGO utilizes a Web Ontology Language (OWL)-based data model to represent knowledge related to mechanisms of pathogenesis and observes principles set forth by the Open Biomedical Ontologies (OBO) Foundry, which aims to promote standardization and interoperability for biomedical ontologies [21]. Mechanisms of pathogenesis are specified where a mechanism is considered to be "the means by which an effect is produced or brought about". The term space is organized in two branches that separate direct and indirect mechanisms of pathogenicity. These two terms contain thirteen and eight first level children, respectively, (Fig. 1) that capture the breadth of known mechanisms with depth elaborated through expansion of subclass hierarchies.

The gene ontology and biological process terms detailing pathogenic mechanisms

The Gene Ontology is a data framework representing biological systems, ranging from the molecular to the organism level, in a species-agnostic manner. As of November 2024, GO contains over 40,000 concepts encompassing signaling and metabolic pathways, developmental processes, cell cycle, etc. (https://release.geneontology.org/2024-11-03/index.html). GO has three aspects which form axes that should provide a complete picture of the function of a gene:

1) Molecular Functions represent activities performed by gene products, such as "catalysis" or "transcription regulator activity"
2) Biological Processes signify 'biological programs' accomplished by the concerted action of multiple molecular functions
3) Cellular Components denote the cellular location in which the molecular function of the gene product occurs

For Sequences of Concern, the main relevant branch of GO is GO:0044003 symbiont-mediated perturbation of host process. Most of the developed GO terms are shown in Tables 1, 2, 3, 4, 5, 6, 7, and 8. Molecularly, the function of the SOC can be any GO Molecular Function, since these are not restricted to interspecies interactions: for example, GO:0016248 channel inhibitor activity, GO:0010856 adenylate cyclase activator activity, GO:0090729 toxin activity. Since these are orthogonal to the BP aspect or a gene product's function, any one of

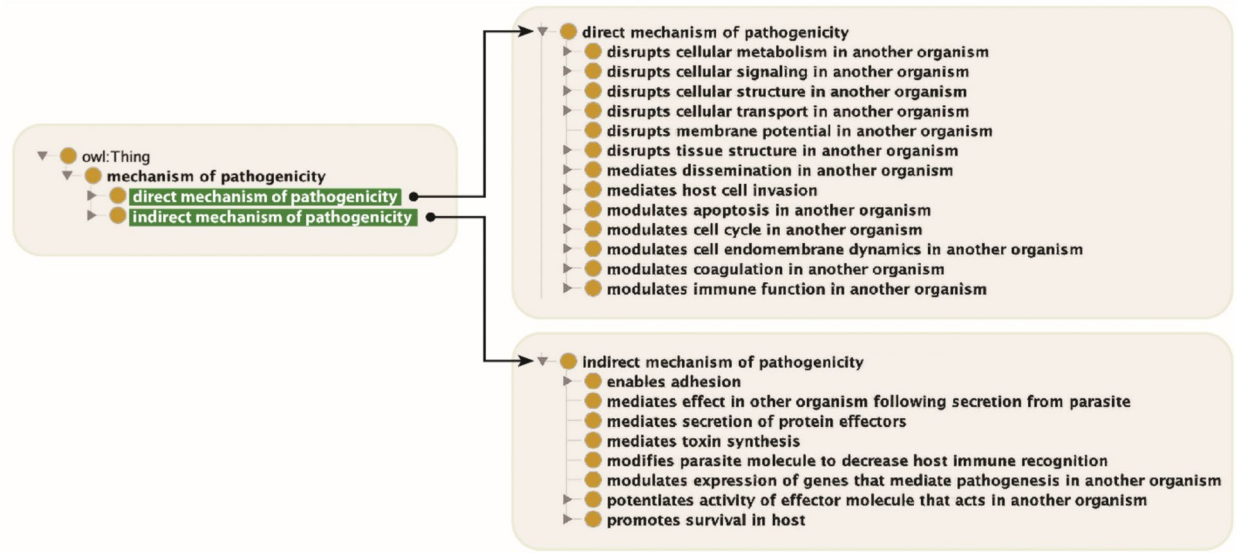

Fig. 1 Tree view of PathGO terms. Left: Top level Structure. Right: First level substructure in direct and indirect mechanism branches

Table 1 Damage to host from cytotoxicity and cell permeabilization

GO ID	GO Term
GO:0001897	symbiont-mediated cytolysis of host cell
GO:0001907	symbiont-mediated killing of host cell
GO:0044658	pore formation in host by symbiont
GO:0141130	symbiont-mediated inactivation of host ribosome
GO:0141042	symbiont-mediated cAMP intoxication of host cell
GO:0052009	symbiont-mediated disruption of host cell wall

Table 2 Damage to host tissue

GO:0141139	symbiont-mediated disruption of host mucosa
GO:0141146	symbiont-mediated disruption of host tissue
GO:0141023	symbiont-mediated disruption of host cell–cell adhesion
GO:0141041	disruption of extracellular matrix of another organism

Table 3 Damage to host from inflammation

GO:0141079	symbiont-mediated activation of host inflammasome-mediated signal transduction
GO:0085033	symbiont-mediated activation of host NF-κB cascade
GO:0141071	symbiont-mediated activation of host MAPK cascade

these functions can be combined with any BPs from the GO:0044003 symbiont-mediated perturbation of host process GO branch.

GO recently narrowed and clarified the definition of a biological process (one of the 3 GO 'aspects [21]; see Gaudet et al., 2017 for more details about the organization of the GO) [22], as *"the execution of a genetically-encoded biological module or program. It consists of all the steps required to achieve the specific biological objective of the module. A biological process is accomplished by a particular set of molecular functions carried out by specific gene products (or macromolecular complexes), often in a highly regulated manner and in a particular temporal sequence."* [23] This led to the removal (obsoletion) of a number of terms that were groupings of similar phenotypes rather than biological programs, such as some signaling or developmental pathways. One of these terms was "pathogenesis" (GO:0009405), which was removed in 2021. At the time of its obsoletion, over 277,000 UniProt accession numbers were annotated with the term [7]. The overwhelming number of cases resulted from automated annotations of UniProt entries containing the UniProt keyword "Pathogenesis". The term was out of scope for GO since pathogenesis does not describe a set of coordinated activities leading to the execution of a biological program. The annotated genes were a mixed bag of sequences that either interfered with host function or affected the symbiont's fitness. Moreover, because the term was so broad, it was impossible to assess from these annotations how the sequence was involved in pathogenesis, what was the targeted host sequence or system, and what was the targeted host taxon. Finally, this term had no children that could have provided more pertinent information.

Table 4 Immune evasion, passive

GO ID	GO Term
GO:0039699	symbiont-mediated evasion of mRNA degradation by host via mRNA cap methylation
GO:0141141	symbiont-mediated evasion of recognition by host pattern recognition receptor
GO:0141178	symbiont-mediated evasion of recognition by host complement
GO:0141179	symbiont-mediated evasion of recognition by host antimicrobial peptide

Table 5 Subversion of host immune signaling

GO ID	GO Term
GO:0141105	symbiont-mediated suppression of host toll-like receptor signal transduction
GO:0039537	symbiont-mediated suppression of cytoplasmic pattern recognition receptor signaling pathway
GO:0039723	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of TBK1 activity
GO:0039724	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of IKBKE activity
GO:0140886	symbiont-mediated suppression of host interferon-mediated signaling pathway
GO:0141080	symbiont-mediated activation of host interferon signaling pathway
GO:0039502	symbiont-mediated suppression of host type I interferon-mediated signaling pathway
GO:0140884	symbiont-mediated suppression of host type II interferon-mediated signaling pathway
GO:0140885	symbiont-mediated suppression of host type III interferon-mediated signaling pathway
GO:0039548	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of IRF3 activity
GO:0039557	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of IRF7 activity
GO:0039514	symbiont-mediated suppression of host JAK-STAT cascade
GO:0039560	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of host IRF9 activity
GO:0039574	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of host TYK2 activity
GO:0039576	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of JAK1 activity
GO:0039562	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of STAT activity
GO:0039563	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of STAT1 activity
GO:0039564	symbiont-mediated suppression of host JAK-STAT cascade via inhibition of STAT2 activity
GO:0141074	symbiont-mediated suppression of host cGAS-STING signal transduction
GO:0141078	symbiont-mediated suppression of host RIG-I signaling pathway
GO:0039540	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of RIG-I activity
GO:0039554	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of MDA-5 activity
GO:0039545	symbiont-mediated suppression of host cytoplasmic pattern recognition receptor signaling pathway via inhibition of MAVS activity)
GO:0039580	symbiont-mediated suppression of host PKR/eIFα signaling
GO:0085032	symbiont-mediated perturbation of host NF-κB cascade
GO:0085034	symbiont-mediated suppression of host NF-κB cascade
GO:0052080	symbiont-mediated perturbation of host MAPK cascade
GO:0141070	symbiont-mediated suppression of host MAPK cascade
GO:0141072	symbiont-mediated suppression of host tumor necrosis factor signaling pathway
GO:0039527	symbiont-mediated suppression of host TRAF-mediated signal transduction
GO:0039579	symbiont-mediated suppression of host ISG15-protein conjugation
GO:0141135	symbiont-mediated suppression of host chemokine signal transduction pathway
GO:0141081	symbiont-mediated suppression of host inflammasome-mediated signal transduction
GO:0141083	symbiont-mediated suppression of host reactive oxygen species generation

For broader adoption of the knowledge gained during sequence annotation for the IARPA Fun GCAT project as reflected in the terms that were developed (FunSoCs and PathGO), it was decided to revise the gene ontology (GO) biological process terms under GO:0044419: "biological process involved in interspecies interaction between organisms". GO is widely used and interoperable with other resources and ontologies. PathGO was limited

Table 6 Subversion of host immune effectors

GO ID	GO Term
GO:0141140	symbiont-mediated suppression of host immunoglobulin-mediated immune response
GO:0141114	symbiont-mediated suppression of host complement activation
GO:0141115	suppression of complement activation by another organism by inactivation of complement proteins
GO:0141116	suppression of complement activation by another organism by complement sequestering
GO:0141117	symbiont-mediated suppression of host complement activation by recruitment of complement control protein
GO:0141203	symbiont-mediated suppression of host complement activation by recruitment of host proteases
GO:0052067	symbiont-mediated perturbation of host phagocytosis
GO:0141073	symbiont-mediated perturbation of host opsonization
GO:0141145	symbiont-mediated suppression of host neutrophil extracellular trap formation
GO:0140133	symbiont-mediated suppression of host cytokine production
GO:0141173	symbiont-mediated suppression of host pro-inflammatory cytokine signaling
GO:0141174	symbiont-mediated suppression of host anti-inflammatory cytokine signaling
GO:0141184	symbiont-mediated activation of host anti-inflammatory cytokine signaling
GO:0039588	symbiont-mediated suppression of host antigen processing and presentation
GO:0046776	symbiont-mediated suppression of host antigen processing and presentation of peptide antigen via MHC class I
GO:0039505	symbiont-mediated suppression of host antigen processing and presentation of peptide antigen via MHC class II
GO:0141059	symbiont-mediated disruption of host antimicrobial peptide activity
GO:0141082	symbiont-mediated detoxification of host-generated reactive oxygen species

Table 7 Host invasion, adherence to host, dissemination in host

GO ID	GO Term
GO:0044409	entry into host
GO:0044417	translocation of molecules into host
GO:0141018	adhesion of symbiont to host via host extracellular matrix
GO:0141024	adhesion of symbiont to host cell surface via host membrane carbohydrate
GO:0141025	adhesion of symbiont to host cell surface via host glycoprotein
GO:0141026	adhesion of symbiont to host cell surface via host membrane cholesterol
GO:0035756	symbiont-mediated migration across host epithelium
GO:0044067	symbiont-mediated perturbation of host cell–cell junction
GO:0106259	symbiont mediated cell-to-cell migration in host
GO:0141142	symbiont-mediated migration across host tissue barrier

in that it mixed molecular function with biological process terms so that they could not readily be imported into GO. Instead of attempting to revise PathGO, it was deemed more practical to transfer the relevant terms to GO.

GO aims to describe the normal, evolved function of genes. In the case of genes from pathogenic microbes that adversely affect the host, these have apparently evolved to exert these effects. They include suppressing immunity and overcoming barriers for the sake of microbial colonization, replication, and transmission. Secondarily, these effects produce a loss of homeostasis (=disease) in the host.

GO has been described as "under-utilized for prokaryotes, single-celled eukaryotic species, and viruses" when compared to model multi-cellular eukaryote species [24]. Expanding the terms relating to pathogenic microbes for GO should be useful for systems biologists and data scientists interested in studying the biological processes of infectious diseases. The Gene Ontology (GO) project offers a structured network of interconnected 'terms' or 'classes' that define the functions of gene products. It also explicitly links these terms to the corresponding gene products that perform these functions. This framework is particularly tailored to facilitate the computational modeling of biological systems across all organisms [21, 25].

Table 8 Manipulation of host cell biology

GO ID	GO Term
GO:0052026	symbiont-mediated perturbation of host transcription
GO:0039604	symbiont-mediated suppression of host translation
GO:0141154	symbiont-mediated suppression of host-directed shutoff of host translation
GO:0039606	symbiont-mediated suppression of host translation initiation
GO:0141155	symbiont-mediated suppression of host translation elongation
GO:0052042	symbiont-mediated activation of host programmed cell death
GO:0052150	symbiont-mediated perturbation of host apoptosis
GO:0033668	symbiont-mediated suppression of host apoptosis
GO:0052151	symbiont-mediated activation of host apoptosis
GO:0044071	symbiont-mediated perturbation of host cell cycle progression
GO:0039648	symbiont-mediated perturbation of host ubiquitin-like protein modification
GO:0044082	symbiont-mediated perturbation of host small GTPase-mediated signal transduction
GO:0141127	symbiont-mediated perturbation of host Rab *small GTPase* signal transduction
GO:0044083	symbiont-mediated perturbation of host Rho small GTPase signal transduction
GO:0052025	symbiont-mediated perturbation of host cell endomembrane system
GO:0044075	symbiont-mediated perturbation of host vacuole organization
GO:0141157	symbiont-mediated suppression of host exocytosis
GO:0141158	symbiont-mediated suppression of host phagosome maturation
GO:0141159	symbiont-mediated suppression of host phagosome acidification
GO:0141160	symbiont-mediated disruption of host phagosome
GO:1,990,215	symbiont-mediated perturbation of host vesicle-mediated transport
GO:0141213	symbiont-mediated generation of symbiont-containing vacuole
GO:0052039	symbiont-mediated perturbation of host cytoskeleton
GO:0141028	symbiont-mediated perturbation of host microtubule cytoskeleton
GO:0141029	symbiont-mediated disruption of host focal adhesion
GO:0141030	symbiont-mediated perturbation of host actin cytoskeleton via filamentous actin depolymerization
GO:0141031	symbiont-mediated perturbation of host actin cytoskeleton via actin crosslinking
GO:0141032	symbiont-mediated perturbation of host actin cytoskeleton via actin filament reorganization
GO:0141033	symbiont-mediated perturbation of host actin cytoskeleton via actin polymerization
GO:0141034	symbiont-mediated perturbation of host actin cytoskeleton via inhibition of actin polymerization
GO:0075071	symbiont-mediated perturbation of host autophagy
GO:0140321	symbiont-mediated suppression of host autophagy

Annotating pathogenic functions with GO terms

This revision improved GO term structure, definitions, and consistency in the biological process involved in the interspecies interaction branch (GO:0044419). During the course of the revision, we eliminated what we felt were repeated, misleading phrases referring to homeostatic regulation within an organism. We concluded that these phrases, including "negative regulation" and "positive regulation" should be avoided when describing multiorganism species interactions because the language of regulation is inappropriate for pathogenic interactions, though it could be appropriate for mutualistic or commensal interactions. For pathogenesis, the goals of the symbiont and the host are largely in conflict, with the pathogenic symbiont attempting to either subvert or evade the normal, evolved operations of the host immune system while the host attempts to limit pathogen spread within it, sometimes even suffering damage from its own response to the pathogen.

Terms describing multi-organism processes should not obfuscate which organism has the initiative when one of the two is responsible for the activity. To show this, we have resorted to the syntax of "symbiont-mediated" to indicate that a symbiont sequence is the instigator. Previously developed terms such as "viral entry into host cell" (GO:0046718) were written as agnostic so as to "annotate both viral and host proteins participating in the entry process" [24].

In the course of renovating GO, we have begun stripping from the term the type of microbe involved in the interspecies interaction (viral, bacterial, protozoal,

fungal). The taxon of the sequence involved in the function described by the GO term can be determined directly from the accession number of the sequences specified. Moreover, making the terms more universal will be useful for identifying commonalities between microbial pathogens.

In authoring new terms for subversion of innate immune signaling, we tried to reference each discrete signaling component for cells that contribute to innate immune defense. We also generated new terms for sequences that (1) alter host cytoskeletal dynamics, (2) enable 'passive' immune evasion by altering microbial molecules so they are less detectable, (3) frustrate host complement activity, (4) adhere to different types of host cell surface molecules, and (5) change host endomembrane biology. We expanded terms to recognize different host molecules to which microbial adhesins and attachment proteins adhere (and attach), both on host cell and within the extracellular matrix.

The syntax can be generalized as:

"symbiont-mediated (perturbation/ suppression/ activation) of host [*biological process*]"

These new terms capture discrete ways in which sequences from microbial pathogens exploit specific host processes so that both machines and humans can better recognize them. Many of the terms are listed in the tables in the following section where they are correlated with the relevant FunSoC term. In addition, we include a supplementary spreadsheet of 320 proteins from ~ 120 species (bacteria, viruses, protozoa, fungi, and a parasitic fluke) each with a UniProt accession, and annotated with 95 of the terms, illustrating their use (Supplemental_JBS_GO-PathGO_annotations_2.xlsx). GO can be downloaded at https://geneontology.org/docs/download-ontology/ and browsed at https://amigo.geneontology.org/amigo. We hope these new and renovated GO terms will lead to general improvements in SoC annotation in secondary and composite databases. We anticipate this will allow bioinformaticians, systems biologists, and other biological data scientists to investigate commonalities across a range of hosts and symbionts.

New and renovated symbiont-host GO terms

In the following eight tables, 95 new and revised GO terms that are children to "biological process involved in interspecies interaction between organisms" (GO:0044419) and directly relevant to microbial pathogenesis are presented. Tables 1, 2, and 3 detail ways in which a microbe can damage a host. Table 4 lists processes relevant to how a microbe can evade the host innate immune detection. Table 5 describes processes by which a microbe can actively subvert host innate immune signaling. Table 6 contains processes by which the symbiont frustrates host innate immune effectors downstream of signaling. Table 7 lists terms related to attachment (adherence), invasion, and dissemination in a host. Table 8 describes some of the ways in which a parasitic symbiont can manipulate host cell biology.

Conclusion

To better understand and report on sequences of concern, we have improved controlled vocabularies required for an adequate description of interspecies pathogenic interactions generating deleterious consequences for the host. The set of GO terms we developed are appropriate for describing the mechanisms of sequences that pathogens regularly deploy to infect hosts, whether animal or plant. These terms enable the commonalities of the struggle of pathogens with hosts to be captured and understood. These include sequences used by pathogens to overcome host barriers, subvert innate immune systems, disseminate within hosts, and otherwise negatively affect critical components of host organisms.

These new terms, specific to interspecies interactions particular to microbial pathogenesis, provide a consistent and defensible (when properly referenced) categorization for sequences of concern. These terms are useful for the systems biology of infectious diseases and can serve as warning triggers for synthetic nucleic acid screening under the new framework. If applied to a "broad enough" set of sequences, then they might also be helpful for the categorization of research for departmental review under the 2024 OSTP United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential that will be in effect in May 2025 [26]. These ontologies provide a controlled but extensive vocabulary that can instantiate the more abstract operative terms in these policies. In particular they clarify and resolve many of the ambiguities in the terms "pathogenicity" and "toxicity" for microbes and other organisms that can threaten human and animal health. We hope they will also enable a better understanding of anti-host microbial processes for infectious disease data science in such a way that could lead to better and more general countermeasures against ID. More and better annotations of these sequences is needed as the United States moves to a biosecurity regulatory framework that considers more than simply the taxonomy of problematic microbes but how the sequences mediating pathogenesis contribute to that state. The work presented here aims to support these efforts.

Abbreviations

DHHS	Department of Health and Human Services
Fun GCAT	Functional Genomic and Computational Assessment of Threats
FunSoCs	Function of Sequences of Concern
GO	Gene Ontology
IARPA	Intelligence Advanced Research Project Activity

OSTP	White House Office of Science and Technology Policy
OWL	Web Ontology Language
PathGO	Pathogenesis Gene Ontology
SoC	Sequence of Concern

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13326-025-00323-8.

Supplementary Material 1.

Acknowledgements
The FunSoC acronym originated with our collaborator and Fun GCAT co-PI, Todd Treangen, a professor of computer science at Rice University in Houston, TX. We are grateful to Dr. C. Matthew Sharkey for reviewing the manuscript and making many helpful comments.
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 U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Authors' contributions
All three authors (G.G., J.P., P.G.) were responsibles for conceptualization, data curation, formal analysis, and investigation of the topic as well as writing the original draft of the manuscript. Figure 1 was provided by J.P. The tables and the supplemental material were constructed by G.G. The manuscript was reviewed and edited by G.G. and P.G.

Funding
Some of the work described herein was the result of support received from the Fun GCAT 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. The Gene Ontology Consortium is funded by the National Human Genome Research Institute (US National Institutes of Health), grant number HG012212, with co-funding by NIGMS.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Signature Science, LLC, 1670 Discovery Drive, Charlottesville, VA 22911, USA. [2]Asymmetric Operations Sector, The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA. [3]SIB Swiss Institute of Bioinformatics and GO Central, 4 rue Michel-Servet, Geneva 1211, Switzerland.

Received: 7 November 2024 Accepted: 25 February 2025
Published online: 21 March 2025

References

1. Office of Science and Technology Policy (OSTP). Framework for nucleic acid synthesis screening. 2024.
2. DHHS A. Screening framework guidance for providers and users of synthetic nucleic acids. 2023.
3. IARPA. Functional genomic and computational assessment of threats. 2016. Available from: https://www.iarpa.gov/research-programs/fun-gcat. Last accessed: 1/29/2024.
4. Ovi F, Zhang L, Nabors H, et al. A compilation of virulence-associated genes that are frequently reported in avian pathogenic Escherichia coli (APEC) compared to other E. coli. J Appl Microbiol. 2023;134(3):lxad014. https://doi.org/10.1093/jambio/lxad014.
5. Olson PD, Hunstad DA. Subversion of host innate immunity by uropathogenic escherichia coli. Pathogens. 2016;5(1):2. https://doi.org/10.3390/pathogens5010002.
6. Wassenaar TM, Gaastra W. Bacterial virulence: can we draw the line? FEMS Microbiol Lett. 2001;201(1):1–7. https://doi.org/10.1111/j.1574-6968.2001.tb10724.x.
7. Godbold GD, Kappell AD, LeSassier DS, et al. Categorizing sequences of concern by function to better assess mechanisms of microbial pathogenesis. Infect Immun. 2022;90(5):e0033421. https://doi.org/10.1128/IAI.00334-21.
8. Balaji A, Kille B, Kappell AD, et al. SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. Genome Biol. 2022;23(1):133. https://doi.org/10.1186/s13059-022-02695-x.
9. Jacak R, Godbold G, Ernlund A, et al. PathGO: the pathogenesis gene ontology. 2021.
10. Gemler BT, Mukherjee C, Howland CA, et al. Function-based classification of hazardous biological sequences: demonstration of a new paradigm for biohazard assessments. Front Bioeng Biotechnol. 2022;10:979497. https://doi.org/10.3389/fbioe.2022.979497.
11. Godbold GD, Scholz MB. Annotation of functions of sequences of concern and its relevance to the new biosecurity regulatory framework in the United States. Appl Biosafety. 2024:1–8; https://doi.org/10.1089/apb.2023.0030.
12. Pei G, Dorhoi A. NOD-like receptors: guards of cellular homeostasis perturbation during infection. Int J Mol Sci. 2021;22(13):6714. https://doi.org/10.3390/ijms22136714.
13. Fitzgerald KA, Kagan JC. Toll-like receptors and the control of immunity. Cell. 2020;180(6):1044–66. https://doi.org/10.1016/j.cell.2020.02.041.
14. Danelishvili L, Everman J, Bermudez LE. Mycobacterium tuberculosis PPE68 and Rv2626c genes contribute to the host cell necrosis and bacterial escape from macrophages. Virulence. 2016;7(1):23–32. https://doi.org/10.1080/21505594.2015.1102832.
15. Chui AJ, Okondo MC, Rao SD, et al. N-terminal degradation activates the NLRP1B inflammasome. Science. 2019;364(6435):82–5. https://doi.org/10.1126/science.aau1208.
16. Bose S, Segovia JA, Somarajan SR, et al. ADP-ribosylation of NLRP3 by Mycoplasma pneumoniae CARDS toxin regulates inflammasome activity. mBio. 2014;5(6):e02186–14; https://doi.org/10.1128/mBio.02186-14.
17. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform. 2015;16(6):1069–80. https://doi.org/10.1093/bib/bbv011.
18. Torto-Alalibo T, Collmer CW, Gwinn-Giglio M. The Plant-Associated Microbe Gene Ontology (PAMGO) consortium: community development of new gene ontology terms describing biological processes involved in microbe-host interactions. BMC Microbiol. 2009;9 Suppl 1(Suppl 1):S1. https://doi.org/10.1186/1471-2180-9-S1-S1.
19. Giglio MG, Collmer CW, Lomax J, et al. Applying the gene ontology in microbial annotation. Trends Microbiol. 2009;17(7):262–8. https://doi.org/10.1016/j.tim.2009.04.003.
20. Korves T, Colosimo ME. Controlled vocabularies for microbial virulence factors. Trends Microbiol. 2009;17(7):279–85. https://doi.org/10.1016/j.tim.2009.04.002.
21. Gene Ontology Consortium, Aleksander SA, Balhoff J, et al. The gene ontology knowledgebase in 2023. Genetics. 2023;224(1):iyad031. https://doi.org/10.1093/genetics/iyad031.
22. Gaudet P, Škunca N, Hu JC, et al. Primer on the gene ontology. Methods Mol Biol. 2017;1446:25–37. https://doi.org/10.1007/978-1-4939-3743-1_3.
23. Gene Ontology Consortium. Gene ontology, 2024_06–17 release. 2017.
24. Foulger RE, Osumi-Sutherland D, McIntosh BK, et al. Representing virus-host interactions and other multi-organism processes in the Gene Ontology. BMC Microbiol. 2015;15:146. https://doi.org/10.1186/s12866-015-0481-x.
25. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
26. Office of Science and Technology Policy (OSTP). United States Government policy for oversight of dual use research of concern and pathogens with enhanced pandemic potential. 2024.

Publisher's Note