frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 07 October 2022
DOI 10.3389/fbioe.2022.979497

OPEN ACCESS

EDITED BY
Patricia Machado Bueno Fernandes,
Federal University of Espirito Santo,
Brazil

REVIEWED BY
Junjie Yue,
Beijing institute of Biotechnology, China
David Gillum,
Arizona State University, United States
Rebecca Mackelprang,
Engineering Biology Research
Consortium, United States

*CORRESPONDENCE
Craig M. Bartling,
bartlingc@battelle.org

SPECIALTY SECTION
This article was submitted to Biosafety and Biosecurity,
a section of the journal
Frontiers in Bioengineering and Biotechnology

RECEIVED 27 June 2022
ACCEPTED 31 August 2022
PUBLISHED 07 October 2022

CITATION
Gemler BT, Mukherjee C, Howland CA, Huk D, Shank Z, Harbo LJ, Tabbaa OP and Bartling CM (2022), Function-based classification of hazardous biological sequences: Demonstration of a new paradigm for biohazard assessments. *Front. Bioeng. Biotechnol.* 10:979497. doi: 10.3389/fbioe.2022.979497

Function-based classification of hazardous biological sequences: Demonstration of a new paradigm for biohazard assessments

Bryan T. Gemler, Chiranjit Mukherjee, Carrie A. Howland, Danielle Huk, Zachary Shank, Lela Johnson Harbo, Omar P. Tabbaa and Craig M. Bartling*

Battelle Memorial Institute, Columbus, OH, United States

Bioengineering applies analytical and engineering principles to identify functional biological building blocks for biotechnology applications. While these building blocks are leveraged to improve the human condition, the lack of simplistic, machine-readable definition of biohazards at the function level is creating a gap for biosafety practices. More specifically, traditional safety practices focus on the biohazards of known pathogens at the organism-level and may not accurately consider novel biodesigns with engineered functionalities at the genetic component-level. This gap is motivating the need for a paradigm shift from organism-centric procedures to function-centric biohazard identification and classification practices. To address this challenge, we present a novel methodology for classifying biohazards at the individual sequence level, which we then compiled to distinguish the biohazardous property of pathogenicity at the whole genome level. Our methodology is rooted in compilation of hazardous functions, defined as a set of sequences and associated metadata that describe coarse-level functions associated with pathogens (e.g., adherence, immune subversion). We demonstrate that the resulting database can be used to develop hazardous "fingerprints" based on the functional metadata categories. We verified that these hazardous functions are found at higher levels in pathogens compared to non-pathogens, and hierarchical clustering of the fingerprints can distinguish between these two groups. The methodology presented here defines the hazardous functions associated with bioengineering functional building blocks at the sequence level, which provide a foundational framework for classifying biological hazards at the organism level, thus leading to the improvement and standardization of current biosecurity and biosafety practices.

KEYWORDS

biohazard, sequence screening, virulence factor, biosecurity, biosafety

Introduction

The rapidly emerging discipline of bioengineering is enabling practitioners to analyze and assemble biological materials and microorganisms for industrial and research purposes through the creation of modified or novel organisms with specific functionalities (Slusarczyk et al., 2012). Bioengineering leverages sequences inspired from natural organisms that have been identified through studies in the life sciences (Figure 1). Exemplar chassis, such as *Escherichia coli* have been engineered with numerous functions, such as those to sense other bacteria, breakdown biofilms, and release toxic payloads (Hwang et al., 2017). While bioengineering is resulting in great benefit to mankind through medical advancements (e.g., precision medicine) and industrial use, the rapid progression and democratization of biotechnologies have presented new challenges for traditional biosafety and biosecurity practices.[1] Current biosafety practices often focus on organisms at the species level, instead of the functional level, which hinders the ability to predict and accurately prepare for previously uncharacterized organisms, such as biodesigns (i.e., engineered organisms) with novel functionalities. For example, focused by a selected list of pathogens, appropriate laboratory safeguards can be put in place using Biosafety Levels promoted by the Centers for Disease Control and Prevention (CDC), which are based on the severity of the disease and infectivity of the organism being manipulated (U.S. Department of Health and Human Services, 2014). While useful in the current paradigm, these biosafety practices do no enable objective and clear guidelines for engineered organisms outside of prioritized lists of species.

Beyond laboratory safety, frameworks to bolster biosafety practices are in place in some countries for research approval (US Department of Health and Human Services, 2017) and DNA ordering (US Department of Health and Human Services, 2022). Current DNA screening practices used by the International Gene Synthesis Consortium (IGSG) follow a uniform screening protocol against a Restricted Pathogen Database (RPD) "derived from international pathogen and toxin sequence databases" (International Gene Synthesis Consortium, 2017). While practical for regulated pathogens, screening sequences against the RPD has led to high false positive rates and requires time-consuming manual screening. In addition to hazardous pathogens and toxins, current best practices are in place for chemical synthesis and distribution of controlled drugs (U.S. Department of Justice. Drug Enforcement Agency. Diversion Control Division, 2019) and chemical weapons (Headquarters Department of the Army, 2018), but bioengineering is enabling bioproduction of such materials (e.g., (Galanie et al., 2015; Nakagawa et al., 2016)), which may also require extra precaution for laboratory manipulation. Given the exponential rise in DNA synthesis orders (Vickers and Small, 2018) and widespread creation of biodesigns, current screening practices using traditional approaches are unsustainable due to the high cost burden (due to high labor costs associated with reviewing sequences) relative to the increasing low cost of nucleotide synthesis. Thus, the need exists to shift from a subjective, organism-centric to an objective (and cost-effective), function-centric biohazard identification and classification system. This need is at the forefront of best practices as new draft guidance for screening synthetic nucleotide orders opens the aperture for screening to "sequences of concern" from select and non-select agents from all nucleotide sequence types—including short sequences (Federal Registar, 2022).

Here we introduce the term "hazardous function," which refers to one or more sequences (and associated metadata) that are associated with pathogenicity, toxicity, drug production, and other functions as described in this paper. Hazardous functions are driven by proteins that provide the organism or system (in the case of a cell free system or cell factory producing a toxin for example) with the necessary properties to cause infection or other detrimental effects. For example, lethal factor from *Bacillus anthracis* is a hazardous function, whereas DNA polymerase from *B. anthracis* is not. Manipulation of hazardous function sequences (e.g., recombinant production, genome insertion, mutation, etc.), even for legitimate purposes, could lead to the production of novel or enhanced hazardous products. In fact, precedent has shown that genetic manipulation can lead to biodesigns with high pathogenicity (van Der Most et al., 2000; Whitworth et al., 2005; Velmurugan et al., 2007; Bartra et al., 2008; Kurupati et al., 2010; Luo et al., 2010; Tsang et al., 2010), host bioregulation ability (Borzenkov et al., 1993; Borzenkov et al., 1994; Gold et al., 2007), vaccine escape capability (Serpinskii et al., 1996; Jackson et al., 2001; Zhang, 2003; Kerr et al., 2004; Chen et al., 2011), high transmissibility (Herfst et al., 2012), high toxicity (Francis et al., 2000), controlled drug production capability (Galanie et al., 2015; Nakagawa et al., 2016), and species extinction capability (Esvelt et al., 2014).

Hazardous functions identified through comparative genomic techniques (Gilmour et al., 2013) and related studies have been cataloged in databases containing virulence factors, toxins, and related other sequences (Supplementary Table S2). However, many of these databases are incomplete, poorly maintained, and/or do not have valuable metadata for objective biosafety assessments. Specifically, we and others have found that many of the entries in these databases simply tag sequences as "virulence factors" if attenuation of the activity

1 For this manuscript, the term biosafety refers to practices associated with protecting researchers from biological hazards associated with an organism based on its characteristics (e.g., practices associated with Biosafety Level 3 organisms). The term biosecurity refers the security of biological materials, including ordering of synthetic nucleotides. Thus, understanding the hazards associated with single synthetically made sequences can aid in biosecurity assessments (i.e., fulfilling synthetic nucleotide orders), whereas understanding the pathogenicity of an organism being manipulated in a laboratory can aid in biosafety assessments.

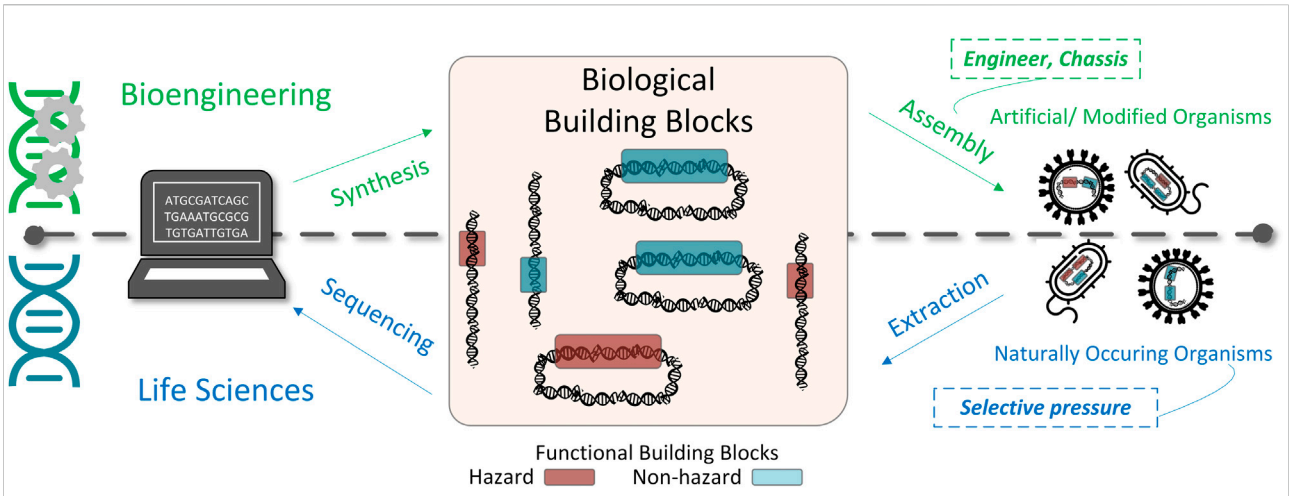

FIGURE 1
Understanding the hazard posed by bioengineering requires the characterization of hazardous building blocks. Bioengineering enables the creation of novel or modified organisms through oligonucleotide synthesis and assembly of building blocks within a biological vehicle. These biological functional building blocks are inspired from sequences extracted and sequenced from natural organisms. Such organisms can be composed of hazardous and non-hazardous functional elements (denoted by the red and blue blocks, respectively). The hazardous functions within organisms may have arisen from environmental selective pressures such as virus mutation to enable, for example, a jump from a vector to a human host, or a transfer of transposable elements among species.

leads to reduced virulence. Thus, many "virulence factors" may not be particularly hazardous in the context of bioengineering. For example, the Victor's Virulence Factors Database (Sayers et al., 2019) compiles bacterial virulence factors implied from published experimentation, such as large-scale mutational screens that seek to identify attenuated virulence phenotypes. Niu et al. illustrated the controversy associated with the term "virulence factor" by determining that 69% (1,368/1,988) of virulence factors in the Virulence Factor Database (VFDB) (Liu et al., 2019a) were common among pathogens and non-pathogens (Niu et al., 2013). In a more specific example, Segura et al. calls into question the definition of "critical virulence factors" for *Streptococcus suis*, suggesting that more scrutiny is needed before characterizing a strain as virulent based on clinical presentation, animal models testing, or *in vitro* tests (Segura et al., 2017). Taken together, current databases do not serve the purpose needed for biohazard identification necessitating the need for better definition and curation around hazardous functions. Godbold et al. recently described a controlled vocabulary called Functions of Sequences of Concern microbial pathogenesis research for bioinformatic applications (Godbold GD et al., 2021). Here we demonstrate the utility of these types of sequences of concern for understanding biohazards associated with bioengineering functional building blocks.

Regardless of the controversy associated with the term *virulence factor*, it is clear that different functions (and context) have different levels of importance for determining the sequence's overall hazard level and thus contribution to the organism or system's hazard level. Given such wealth of publicly available knowledge on the functions derived from genetic sequences in UniProt (and related databases), databases such as those presented in Supplementary Table S2, and the scientific literature at large, the scientific community is primed to enable function-based DNA sequence assessment to aid in the preparation for novel pathogens and/or components with hazardous properties as well as prevent nefarious development of novel engineered pathogens. To anticipate potential hazards associated with novel pathogens, Colf et al. called for "functionality-based approach" that focuses on key hazard elements such as stability of an organism, infectious dose, or toxicity (Colf, 2016), but such practices have not fully come to fruition. Here we introduce a paradigm of function-based sequence assessment that may fill the gaps associated with current biosafety practices. Hazardous functions can be subjective based on what the user considers a "hazard," but here we focus on functions associated with pathogenicity, toxicity, drug production, and other functions that can harm humans or other organisms of interest (e.g., livestock, crops, etc.). We first demonstrate our novel methodology to create a database of hazardous sequences classified into coarse functional categories. We then validate our methodology by demonstrating that a subset of the resulting database can be used to successfully distinguish pathogenic from nonpathogenic organisms via specific functional mechanisms. Finally, we further demonstrate the application of this methodology and resultant database through an example hazard scale. Therefore, the

methodology demonstrated here can immediately be used for biosecurity screening assessments of synthetic genes (through the exemplar hazard scale) and partial biosafety assessments for classification of bacterial pathogens and non-pathogens. Because our methods rely on the DNA sequence's encoded function, rather than agent-based lists, we provide a foundation for enabling function-based hazard assessments. This foundation can be built upon to provide comprehensive biosecurity and biosafety assessments for any novel biodesign through only analysis of the biodesign's genome.

Results

A methodology and database for function-based hazard assessments

To enable function-based biohazard screening, we developed an access-controlled biological Functional Hazards Database that contains protein sequences with metadata. The database documents sequences that have been verified in the laboratory to encode a hazardous function based on experimental information from the primary literature and/or publicly available databases (e.g., Supplementary Table S2). We have compiled these sequences and metadata into a machine-readable database that is focused on biohazards that target humans and non-humans of high economic value. Non-human hosts are based on an analysis performed by the United States Department of Agriculture Economic Research that demonstrated cattle, poultry, and swine comprised 96% of U.S. livestock farm receipts (of $176 billion) and corn, soybeans, and wheat comprised 48% of U.S. crop farm receipts (of 195.4 billion) (United States Department of Agriculture Economic Research Service, 2022) in 2017. Together, these six commodities comprise 71% of all U.S. farm receipts in 2017 (United States Department of Agriculture Economic Research Service, 2022).

We focus our database on particularly hazardous functions, which includes only a subset of virulence factor types as well as several hazardous functions not considered virulence factors (Figure 2). We delineate a virulence factor from a hazardous function as follows: while a virulence factor describes any factor (protein or otherwise) that aids in the virulence of organism, we define functional hazards as any sequence whose *verified* encoded *function* can lead to a direct and harmful impact on a host given a biological vehicle to do so. Thus, a logical division between hazardous functions and virulence factors (Figure 2) emerges based on this definition. Some traditional virulence factors are thus considered hazards, such as those involved in evading the host's immune system which–when encoded in an appropriate biological context (e.g., in *E. coli*)—contribute to direct detrimental impact to the host. In contrast, a transcription factor, for example, may only indirectly impact pathogenicity, and is thus not included in our hazard definition. We further delineate factors that are found throughout nature (i.e., those that are typically not unique to pathogens), such as siderophores, some secretion systems, and some non-protein virulence factor biosynthesis enzymes. For example, Type I and Type II secretion system proteins, which are ubiquitous throughout all gram-negative bacteria—pathogens and non-pathogens (Green and Mecsas, 2016)—are not considered hazardous functions in our definition. In contrast, Types III and IV secretion system proteins, which enable transport of potentially hazardous payloads across two gram-negative bacterial membranes and a host membrane, are considered hazardous functions. Further, careful consideration is given to particularly hazardous non-protein virulence factors such as endotoxin, which is biosynthesized by several enzymes (Raetz and Whitfield, 2002). More importantly, we consider several other sequence types that are not considered traditional virulence factors to be hazardous functions, such as prions, bioregulators, animal toxins (e.g., conotoxins), protein toxins (e.g., ricin), and proteins involved in the biosynthesis of small molecule toxins (e.g., saxitoxin) and drugs (e.g., morphine). For all hazardous sequences, we functionally classify the type of hazardous function into one or more of the 15 high level categories in Table 1 and elaborated below. These categories, chosen based on previous expert discussions from scientists with a variety of life science backgrounds, provide the basis for distinguishing pathogens and nonpathogens as shown by our validation and example biosafety assessment hazard scale discussed later.

Adherence, invasion, and motility

Adherence factors contained within our functional hazard database have experimental evidence (e.g., immunoprecipitation, cell binding assay, etc.) of a direct interaction with host membrane components. Interaction between the adherence factor and the host may enhance host cell tropism through direct interactions of a pathogenic apparatus that binds surface host cell receptors. Proteins that do not directly interact with the host but may be required for assembly of such a pathogenic apparatus can also be considered adherence factors but are further identified in our database as being dependent on direct adherence factors. For example, a type-4 pilus apparatus is responsible for adherence of *Neisseria meningitidis* to host receptors (Rudel et al., 1995), but is composed of several protein subunits. PilC and PilE have direct interactions with the host, whereas other proteins in the assembly do not (Bernard et al., 2014).

Invasion factors are those that leverage mechanisms such as Type III or Type IV Secretion Systems (T3SS/T4SS), pore formation, actin polymerization dysregulation, or cell lysis. The T3SS is a multi-protein needle complex that allows bacterial effectors to be delivered from the pathogen into the host cell directly. These effector proteins promote infection and suppress host defenses. For example, the *Yersinia pestis* T3SS

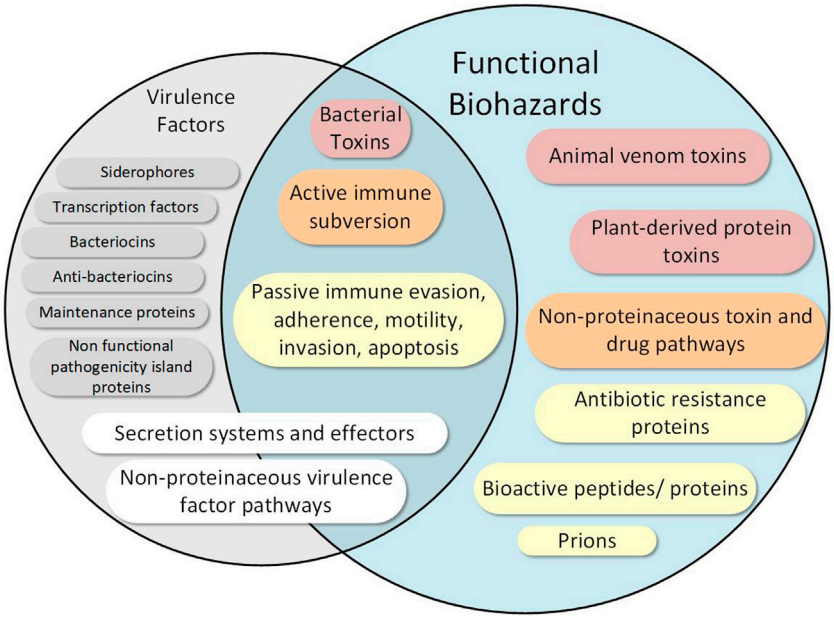

FIGURE 2
Functional biological hazards are differentiated from and include several functions beyond virulence factors. While some hazardous functions overlap with virulence factors, we define several hazardous functions outside the traditional definition of virulence factors. Many virulence factors that may be contained in avirulent organisms, such as siderophores and transcription factors, are not considered hazardous functions as they do not directly and uniquely perform hazardous functions. Hazardous functions are further described in the text and are color coded according coarse functional metadata groups as follows: Red—functions that do direct damage to cells such as toxins; Orange—functions involved in active host subversion or those involved in nonproteinacous toxin and drug pathways; Yellow—other virulence factors uniquely involved in pathogenicity (e.g., invasion), non-virulence factors that may contribute to detrimental host response (e.g., bioregulators and antibiotic resistance proteins), and prions; White—virulence factors that may also participate in non-hazardous microorganism functions; Gray—virulence factors that do not have a direct hazardous function. Note that the figure is non-exhaustive.

TABLE 1 Hazardous functional metadata categories.

Functional metadata	Definition
Adherence	Mediates pathogen or toxin binding to host cell
Motility	Enables a pathogen to move within or between host cells
Invasion	Enables a pathogen or toxin to actively enter or maintain protected spaces within the host
Inhibits host cell death	Inhibits host cell death
Host cell apoptosis	Leads to, aids in, and/or promotes host cell death
Passive host subversion	Passively works to avoid the immune surveillance, e.g., by altering recognizable elements of the pathogen
Active host subversion	Actively aggravates host immune detectors or effectors
Antibiotic resistance	Enables resistance of a pathogen to antibiotics
Damage	Actively damages host cells, host cell processes, or host barriers such as the extracellular matrix. Toxin sequences specifically contain the toxin activity gene ontology term (GO: GO:0009636)
Toxin pathway	Directly involved in the biosynthesis of a non-proteinaceous toxin
Drug pathway	Directly involved in the biosynthesis of a non-proteinaceous drug
Protein Bioregulators	Regulates cellular processes that can be detrimental to the host
Bioregulator pathway	Directly involved in the biosynthesis of a non-proteinaceous bioregulator that can be detrimental to the host
Prion	Protein that can misfold to become an infectious agent
Unknown	Hazardous function is unknown but contributes to complete or near complete loss of virulence when deleted or mutated

structure includes nearly 40 proteins (Cornelis, 2000; Frolkis et al., 2010). In *Y. pestis*, T3SS activation is triggered by cell contact and induces the secretion of effector proteins—termed Yersinia outer proteins (Yops)—across the host cell membrane where they inhibit bacterial phagocytosis and suppress the host immune response (Plano and Schesser, 2013). Like T3SSs, sequences such as bacterial pore-forming lysins and fungal cutinases, which can enable invasion through cleaving host cell walls (Sweigard et al., 1992; Dean et al., 2005; Chen et al., 2007; Basso et al., 2017) are included as well. Other types of invasive bacterial proteins, such as invasion plasmid antigen A (IpaA) from *Shigella sp*., which enables invasion through actin dysregulation (Izard et al., 2006; Park et al., 2011), are also included.

In addition to adherence and invasion, we include some motility factors, as some pathogens use mechanisms that allow a microbe to actively move between or within host cells following infection. This phenomenon is known as actin-based motility, which involves subversion of the host actin cytoskeleton to stimulate movement within the host cell, ultimately leading to microbial spread between cells. This rapid microbial dissemination is a critical step in many infectious diseases. For example, diseases caused by *Listeria monocytogenes* are caused in part by the protein ActA, which directly activates host actin polymerization machinery. This activation results in the formation of an actin "rocket tail" that propels the bacteria into adjacent cells, thereby infecting them (Finlay, 2005; Ireton, 2013).

Host cell death

During infection, pathogens work to maintain tight control of the host's intrinsic cell death mechanisms, often suppressing cell death then activating it to allow replication then dissemination, respectively. Induction of host cell death is used as a pathogenic strategy to allow a virus or bacteria to efficiently exit the host cell, spread to neighboring cells and access nutrients (Ashida et al., 2011). Further, by inducing host cell death, a pathogen can also eliminate immune cells and effectively evade immune defenses (Lamkanfi and Dixit, 2010; Ashida et al., 2011). Viruses are common proponents of this mechanism to facilitate dissemination of replicated virus and suppression of the immune system. For example, the human immunodeficiency virus (HIV), induces programed cell death in healthy T lymphocytes, contributing to the gradual T cell decline and ultimately acquired immune deficiency syndrome (Ahr et al., 2004; Romani and Engelbrecht, 2009). Thus, proteins such those that promote this induction of apoptotic signal (Vpr and HIV envelope proteins) are including in our database (Ayyavoo et al., 1997; Ahr et al., 2004; Romani and Engelbrecht, 2009). In contrast to induction of host cell death, inhibition of host cell death is also a hazardous function since host cell death can be used as an immune defense mechanism to contain the spread of the infection. These hazardous functions enable a pathogen to promote its overall survival within the host by giving the pathogen more time to colonize efficiently prior to dissemination. Enteropathogenic *Escherichia coli*, for example, uses this strategy to stall premature host cell death during infection through the EspZ effector protein, which activates pro-survival signaling pathways within the host (Shames et al., 2010; Shames and Finlay, 2010).

Passive and active host subversion

Pathogens can also evade the host by avoiding or aggregating more specific host immune defenses than those discussed above. Microbes have evolved numerous and diverse strategies to circumvent the host immune system, many even using multiple mechanisms. We classify these strategies as passive or active, in which hazardous functions act to either avoid host immune surveillance or actively interfere with the host's immune responses, respectively. Common passive mechanisms include using antigenic variation, epitope masking, and the use of decoys or molecular mimicry. Often, circumvention of host detection is accomplished by a virulence factor altering recognizable elements of the pathogen. For example, Ebola virus glycoprotein (GP), a key antigen in Ebola pathogenesis, can evade host immune defenses by epitope masking and steric shielding (Cook and Lee, 2013; Wong et al., 2014). Steric shielding of surface epitopes by glycans also prevents antibody binding and binding of host major histone compatible complex I and β1 integrins with other immune cells, thereby preventing the host immune response (Francica et al., 2010). Ebola virus also leverages decoy mechanisms by producing large quantities of secreted GP proteins that adsorb host antibodies (Blair et al., 2015).

In contrast to passive subversion, active host subversion involves active interference with the host's immune responses. For such interference, a microbe must produce factors that are able to block or modulate specific steps in the immune response cascade (Schmid-Hempel, 2009). These factors can be membrane-bound or directly injected directly into the host cell using secretion systems such T3SSs, as discussed above (Raymond et al., 2013). Many bacteria possess efficient means of evading the host complement system. For example, chemotaxis inhibitory protein (CHIPS) from *S. aureus* can bind receptors on neutrophils, blocking their recruitment and engagement to resist complement-mediated killing (Rooijakkers et al., 2005). Active evasion of the immune system can also be accomplished by interfering with the immune response signaling network. For example, *Yersinia* Yop proteins downregulate the expression of TNF-α, thereby effectively blocking pro-inflammatory signaling (Sweet et al., 2007; Schmid-Hempel, 2009).

Antibiotic resistance

Just as pathogens can evade endogenous host responses, pathogens have evolved to evade exogenous factors, such as antibiotics, through expressing hazardous functions.

Surveillance of these hazardous functions is critical, as the rapid and broad dissemination of antibiotic resistance determinants by lateral gene transfer has been demonstrated throughout diverse bacterial species. Several mechanisms have been described that can lead to antibiotic resistance including: production of enzymes capable of metabolizing or modifying antibiotics, antibiotic binding-site modifications to prevent binding, production of outer membrane components that confer low permeability, and overexpression of multi-drug efflux pumps (Fournier et al., 2006; Vila et al., 2007; Kempf and Rolain, 2012; Blair et al., 2015; Bakour et al., 2016; Geisinger and Isberg, 2017). Bacteria often employ more than one mechanism of antibiotic resistance, leading to multidrug-resistant strains. For example, methicillin resistant *S. aureus* (MRSA), produce both β-lactamases that can inactivate β-lactam antibiotics (e.g., penicillin), as well as proteins acquired by lateral gene transfer (PBP2a proteins) that confer resistance to methicillin (Chambers, 1997; Stapleton and Taylor, 2002). While antibiotic resistance factors can be hazardous, the context of the factors needs to be carefully considered. Often antibiotic resistance has been shown to result in virulence attenuation (Andersson and Hughes, 2010; Geisinger and Isberg, 2017), but some studies demonstrate that resistance has increased pathogenic potential during infection (Luo et al., 2005; Skurnik et al., 2013; Roux et al., 2015). While the precise correlation between virulence and antibiotic resistance remains unclear, we define antibiotic resistance as hazardous function given reasonable context (i.e., contained within a pathogen).

Damage

Perhaps the most hazardous functional category can be considered one that does direct damage to the host. While some of the above hazardous functions can directly damage the host, biological toxins represent the largest class of directly damaging hazardous functions. According to the Gene Ontology Consortium, biological toxin activity involves the selective interaction "with one or more biological molecules in another organism (the "target" organism), initiating pathogenesis (leading to an abnormal, generally detrimental state) in the target organism" (EMBL-EBI, 2019). Biological toxins may be proteinaceous or non-proteinaceous, with protein toxins often consisting of multiple subunits that attribute to virulent functions for adherence, invasion, and inactivation of critical cellular functions. Toxins are highly diverse, even within some toxin types. For example, possibly hundreds of thousands of conotoxins—antagonists or agonists of various receptors and ion channels—exist (Lewis et al., 2012). Examples of proteins relevant to this category included in our hazardous function database are shown in Supplementary Table S4.

Pathways

In addition to protein toxins, our database includes key enzymes involved in the biosynthesis of fully and partially characterized small molecule toxin pathways, such as those that produce aflatoxins (cancer-causing and cellular process-disruption fungal toxins (Haschek and Voss, 2013; National Cancer Institute, 2019)), trichothecenes mycotoxins (protein synthesis-inhibiting fungal toxins (Kiessling, 1986)), microcystins (cyanobacterial serine/threonine protein phosphatase-hepatotoxins (Tillett et al., 2000; Campos and Vasconcelos, 2010)), tetrodotoxins (bacterial sodium channel-blocking neurotoxins) (Jal and Khora, 2015; Lago et al., 2015; Magarlamov et al., 2017), and saxitoxins (bacterial sodium channel-blocking neurotoxins) (Al-Tebrineh et al., 2010).

Beyond hazardous pathogens and toxins, we also consider naturally derived or inspired drugs. Bioengineering is presenting a new challenge to control the production of these naturally derived drugs, as the starting materials may not be regulated. Some drugs, such as opiates and cannabinoids, are produced naturally in plants, and have been demonstrated to be produced in yeast and bacteria (Galanie et al., 2015; Poulos and Farnia, 2015; Nakagawa et al., 2016). Illicit drugs pose a hazard to public health and the economy and are thus controlled by the US Drug Enforcement Administration (DEA) using a five category classification system (United States Drug Enforcement Administration, 2019), with schedule I drugs being the highest hazards as they have no currently accepted medical use and have a high potential for abuse (e.g., heroin and cannabis). For chemical synthesis, supplies to synthesize drugs are regulated by the US government (U.S. Department of Justice. Drug Enforcement Agency. Diversion Control Division, 2019), but biosynthetic supplies are less regulated and may thus present a gap in biosecurity and biosafety. Our functional hazards database thus includes exemplar pathways such as the opioid and cannabinoid pathways, which are fairly well elucidated (Galanie et al., 2015; Nakagawa et al., 2016) as well as sequences from less characterized pathways, such as the cocaine pathway (Jirschitzka et al., 2012).

Bioregulators

We also consider host regulators as well, since such molecules can ultimately lead to manifestations of disease (Goldman, 2000) and have drug-like activity. These bioregulators can be peptides, proteins, and small molecules produced naturally by the host in response to an insult or produced by other organisms (e.g., amphibians). Further, regulatory peptides have been discovered and created to mimic small molecule regulators such as opioids (Dudak et al., 2011; Aldrich and McLaughlin, 2012). Like antibiotic resistance factors, the context and scope of bioregulators must be carefully considered. While many bioregulators can be considered hazardous, we limited our initial database to those that could have a high impact on human systems such as the cardiovascular, nervous, and immune systems (Supplementary Table S3).

Prions

Prions are considered a functional hazard as well. A prion is a protein that can misfold to become an infectious agent

(i.e., transmitted from one host to another). Prions most abundantly occur in the brain and are responsible for a variety of fatal progressive neurodegenerative disorders called transmissible spongiform encephalopathies (Prusiner, 1998). The causative agents of these diseases are normal cellular prion proteins (PrPC) that have undergone a posttranslational conformational change into an abnormal scrapie prion protein (PrPSc) (Huang et al., 2015). PrPSc proteins are able to transmit the pathological conformation to PrPC through poorly understood mechanisms (Dobson, 2001; Huang et al., 2015; Erana and Castilla, 2016). Notable prions included in our database are those that lead to Bovine Spongiform Encephalopothy (BSE, or "mad cow disease"), Creutzfeldt-Jakob disease in humans, feline spongiform encephalopathy in cats, and exotic ungulate encephalopathy in zoo animals (Wells et al., 1987; Wilesmith, 1994; Will et al., 1996). Although these diseases are rare, they are usually rapidly progressive and fatal and synthetic versions can induce pathology in experimental animals (Telling et al., 1995; Legname et al., 2004).

Unknown

While many hazardous functions have distinct mechanisms, we do consider potentially hazardous functions with nonspecific mechanisms as well. Throughout the database compilation process, we identified several instances where a protein sequence likely contributes to a hazardous function, but the exact mechanism is unknown. For example, our database contains a relatively high number of Mycobacterium sequences since we leveraged many of the virulence factors documented in PATRIC (Wattam et al., 2017), which relied mainly on one study. In this study, the authors identified which genes are required for *in vivo* growth (and not *in vitro* growth) (Sassetti and Rubin, 2003). Thus, while many of these genes are considered to potentially contribute to hazardous functions, their actual functions are unknown.

Validation of the methodology and resulting functional hazard database: Identification of hazardous functions

To validate our methodology of identifying, categorizing, and databasing hazardous sequences, we leveraged the studies presented in Table 3, which segregate various pathogenic and nonpathogenic bacterial species. We identified eight different organism groups and separated species in each group into pathogens and nonpathogens. We further categorize the pathogens into species and/or disease-causing groups. With the exception of *Pseudomonas syringae* (a plant pathogen), all species leveraged in this validation are pathogenic to humans and/or economically critical livestock. For the validation, we aligned the coding sequences (CDSs) from each strain against a subset of our database that contained only hazardous function sequences from each of the eight organism groups. We used a subset of our database to reduce potential noise associated with hazardous functions potentially encoded in nonpathogens as a proof of concept for the method; thus any use of this methodology for biosafety assessments should note this limitation. We scored each CDS alignment hit as the *(percent identity)* × *(percent hazardous sequence coverage)* and normalized each hit to the total number of CDSs contained in the strain. The normalization step was performed since, for example in the case of *E. coli*, 1 Mb genome size differences can occur among strains, leading to different pathotypes (Dobrindt, 2005). To count the fraction of hazardous CDSs in each strain, we considered different alignment thresholds to ensure that a specific alignment cutoff did not impact our results. Specifically, the fraction of hazardous sequences is nearly unchanged between 20 and 80% alignment scores for all groups (data not shown). Importantly, the fraction of hazardous functions in pathogenic species compared to nonpathogenic species is higher across the entire range in nearly all cases.

Table 2 shows the number and percentage of hazardous CDSs using a relatively stringent alignment threshold of 40%. The 40% threshold has previously been demonstrated to be a useful cutoff by Suzek et al. (2015). In the referenced study, the authors showed 97% of Uniref50 cluster members, defined by the 40% threshold (≥50% sequence identity over 80% sequence coverage (UniProt, 2019a)), share identical or similar gene ontology terms (i.e., have the same function) (Suzek et al., 2015). Thus, this threshold is useful for CDSs that have identical or similar functions relative to sequences contained in the hazardous function database. Table 3 outlines that the average number and fraction of CDSs identified for each pathogenic and nonpathogenic group using the 40% threshold. In 19 out of 21 pathogenic groups, the percentage of CDSs is higher for pathogens compared to nonpathogens (16/18 being significantly higher), suggesting that our methodology was successful in identifying hazardous functions for these groups.

We further identified specific hazardous functions enriched in each pathogenic group (Supplementary Table S1). For this analysis, we assumed (based on testing, data not shown) that a function is "enriched" in a pathogenic group compared to its nonpathogenic counterpart if the average alignment score across all strains in the group is ≥60% higher than the average in the nonpathogen group or the average in the nonpathogen group is 0% and the average in the pathogen group is ≥40%. As a control, we also determined if any hazardous functions are enriched in the nonpathogen group (i.e., if the average alignment score in the nonpathogenic group is ≥60% higher than the pathogen group or the average in the pathogen group is 0% and the average in the nonpathogen group is ≥40%). Based on this analysis, we identified 379 total enriched functions in the pathogenic groups compared to only 12 total hazardous functions in the nonpathogen groups. The pathogen groups averaged 19 enriched hazardous functions across the various pathogen groups (range 1–70, Supplementary Table S1). These functions were involved in a variety of

TABLE 2 The average number and percentage of hazardous CDSs are greater in pathogenic groups compared to nonpathogenic Groups.

Organism	Group	Genera/species in group	Average ± SD of # CDSs with hazardous functions (Average % CDSs)[a]
Neisseria	Pathogenic	*N. meningitidis*	***50 ± 3 (2.3%)***
		N. gonorrhoeae	***53 ± 3 (2.2%)***
	Nonpathogenic	See Table 3	27 ± 5 (1.3%)
Escherichia coli	Pathogenic	EAEC/ETEC/AIEC/EPEC	160 ± 31 (3.2%)
		EHEC	***290 ± 20 (5.3%)***
		ExPEC	***163 ± 32 (3.3%)***
	Nonpathogenic	See Table 3	125 ± 26 (2.7%)
Burkholderia	Pathogenic	*B. mallei*	***111 ± 17 (2.1%)***
		B. pseudomallei	***143 ± 16 (2.1%)***
		B. cenocepacia	***102 ± 8 (1.5%)***
	Nonpathogenic	See Table 3	82 ± 20 (1.2%)
Pseudomonas	Pathogenic	*P. aeruginosa* and *P. mendocina*	***126 ± 21 (2.3%)***
		P. syringae (plant pathogen)	76 ± 3 (1.3%)
	Nonpathogenic	See Table 3	76 ± 7 (1.9%)
Streptococcus	Pathogenic	*S. pneumoniae*	***39 ± 6 (1.9%)***
		S. pyogenes	***42 ± 3 (2.2%)***
		S. suis	***33 ± 4 (1.6%)***
	Nonpathogenic	See Table 3	20 ± 2 (1.0%)
Bacillus	Pathogenic	*B. cereus* and others (See Table 3)	***59 ± 11 (1.1%)***
		B. anthracis	***61 ± 3 (1.1%)***
	Nonpathogenic	See Table 3	23 ± 10 (0.5%)
Clostridium	Pathogenic	*C. botulinum* and *C. tetani*	***6 ± 1 (0.3%)***
		C. difficile	5 ± 1 (0.2%)
		C. perfringens	***5 ± 1 (0.4%)***
	Nonpathogenic	See Table 3	1 ± 1 (0.1%)
Mycobacterium	Pathogenic	*M. tuberculosis* and others (See Table 3)	***440 ± 8 (26%)***
		M. leprae and others (See Table 3)	***281 ± 120 (18%)***
	Nonpathogenic	See Table 3	288 ± 19 (12%)

[a]CDSs above the 40% threshold as defined in the Methods Section; the fraction of CDSs is defined by the number of hits divided by the total number of CDSs in each strain.
Bold italics represents a significant difference in percentage between the pathogenic and nonpathogenic group as defined by a pairwise *t*-test ($p < 0.05$, two-tailed, unequal variance).

processes such as adherence, immune evasion, antibiotic resistance and damage (including toxin activity). The hazardous functions identified to be enriched in the nonpathogen groups mapped to four functions in the *E. coli* group (required for colonization but with unknown mechanisms), one antibiotic resitance function in the *P. syringae* group, three functions in the in the *S. pyogenes* group (involved in antiphagocytosis but with unknown mechanisms), and four antibiotic resistance functions in the Mycobacterium groups. Thus, the results in Supplementary Table S1 suggest that our database enables successful identification of enriched hazardous functions from pathogens as compared to their nonpathogenic counterparts.

TABLE 3 Genomic data from pathogenic and nonpathogenic strains used in this study.

Type	Genera/ Species organism group	References	Pathogenic groups: species/strains (#)	Nonpathogenic groups: species/strains (#)	# Hazardous functions in database
Gram-negative bacteria	Neisseria	Lu et al. (2019)	*1. N. meningitidis* (85) *2. N. gonorrhoeae* (15)	*N. lactamica* (3); *N. longa* (1); *N. zoodegmatis* (1); *N. longate* (1)	67
Gram-negative bacteria	*Escherichia coli*	Cosentino et al. (2013)	1. EAEC/ETEC/AIEC/EPEC (11) 2. EHEC (8) 3. ExPEC (10)	K-12 (2); other non-pathogenic strains (13)	374
Gram-negative bacteria	Burkholderia	Cosentino et al. (2013)	*1. B. mallei* (4) *2. B. pseudomallei* (4) *3. B. cenocepacia* (4)	*B.* sp. *CCGE1001* (1); *B.* sp. *YI23* (1); *B. glumae BGR1* (1); B. *phymatum STM815* (1); *B. phytofirmans PsJN* (1)	141
Gram-negative bacteria	Pseudomonas	Cosentino et al. (2013)	*1. P. aeruginosa* (5) and *P. mendocina* (2) *2. P. syringae* (3)	*P. brassicacearum* (1); *P. fluorescens* (2); *P. putida* (6); *P. stutzeri* (1)	175
Gram-positive bacteria	Streptococcus	Cosentino et al. (2013)	*1. S. pneumoniae* (9) *2. S. pyogenes* (13) *3. S. suis* (9)	*S. parauberis* (1); *S. salivarius* (3); *S. thermophilus* (5)	161
Gram-positive bacteria	Bacillus	Cosentino et al. (2013)	*1. B. cereus* (6); *B. cytotoxicus* (1); *B. weihenstephanensis* (1) *2. B. anthracis* (5)	*B. amyloliquefaciens* (4); *B. atrophaeus* (1); *B. cellulosilyticus* (1); *B. cereus Q1* (1); *B. clausii* (1); *B. coagulans* (2); *B. halodurans* (1); *B. megaterium* (1) *B. pumilus* (1); *B. selenitireducens* (1); *B. subtilis* (4)	116
Gram-positive bacteria	Clostridium	Cosentino et al. (2013)	*1. C. botulinum* (8) and *C. tetani* (1) *2. C. difficile* (2) *3. C. perfringens* (3)	*C. acetobutylicum* (3); *C. beijerinckii* (1); *C. cellulovorans* (1); *C. clariflavum* (1); *C. kluyveri* (2); *C. lentocellum* (1); *C. ljungdahlii* (1); *C. phytofermentans* (1); *C. saccharolyticum* (1); *C.* sp. *SY8519* (1); *C. thermocellum* (1)	54
Bacteria	Mycobacterium	Andreevskaia et al. (2006); Cosentino et al. (2013); Ilina et al. (2013); Prasanna and Mehra (2013)	*1. M. africanum* (1); *M. avium* (1); *M. bovis* (1); *M. canettii* (1); *M. tuberculosis* (5) *2. M. abscessus* (1); *M. avium* (1); *M. leprae* (2); *M. marinum* (1); *M. ulcerans* (1)	*M.* sp. *KMS* (1); *M. gilvum* (1) *M. rhodesiae* (1); *M. smegmatis* (1); *M.* sp. *JLS* (1); *M.* sp. *MCS* (1) *M.* sp. *Spyr1* (1); *M. vanbaalenii* (1)	339

Validation of the methodology and resulting functional hazard database: Hazard fingerprints

To validate the classification component of our methodology (Table 1), we leveraged our functional categories to create "hazard fingerprints" for each strain. The fingerprints were calculated by summing the alignment scores for the CDSs for each strain that belong to each functional category. For these alignments, we accounted for both highly confident hazardous CDSs (e.g., those with alignment scores >40% to our database) as well as less confident, yet potentially hazardous functions by summing all qualified alignment scores as described in the Methods section. This approach allows for more score contribution for higher identity alignments while still allowing for some contribution for lower identity alignments. We then normalized the scores within each functional category by dividing each value by the maximum value in that functional category. This normalization enables critical hazardous functions that may only be encoded with one or a few CDSs (e.g., a critical toxin) that are absent in nonpathogens to be emphasized within a category and controls for abundance bias within our hazard database across functional categories. For this analysis, we considered only known functions (i.e., the "unknown" functional category Table 1 was excluded) to remove noise from the analysis stemming from sequences with potentially hazardous but unknown functionalities. Figure 3 shows the fingerprints for each of the eight organism groups in the form of heat plots to study visual differences among the various hazard categories. We further analyzed the hazard fingerprint data from

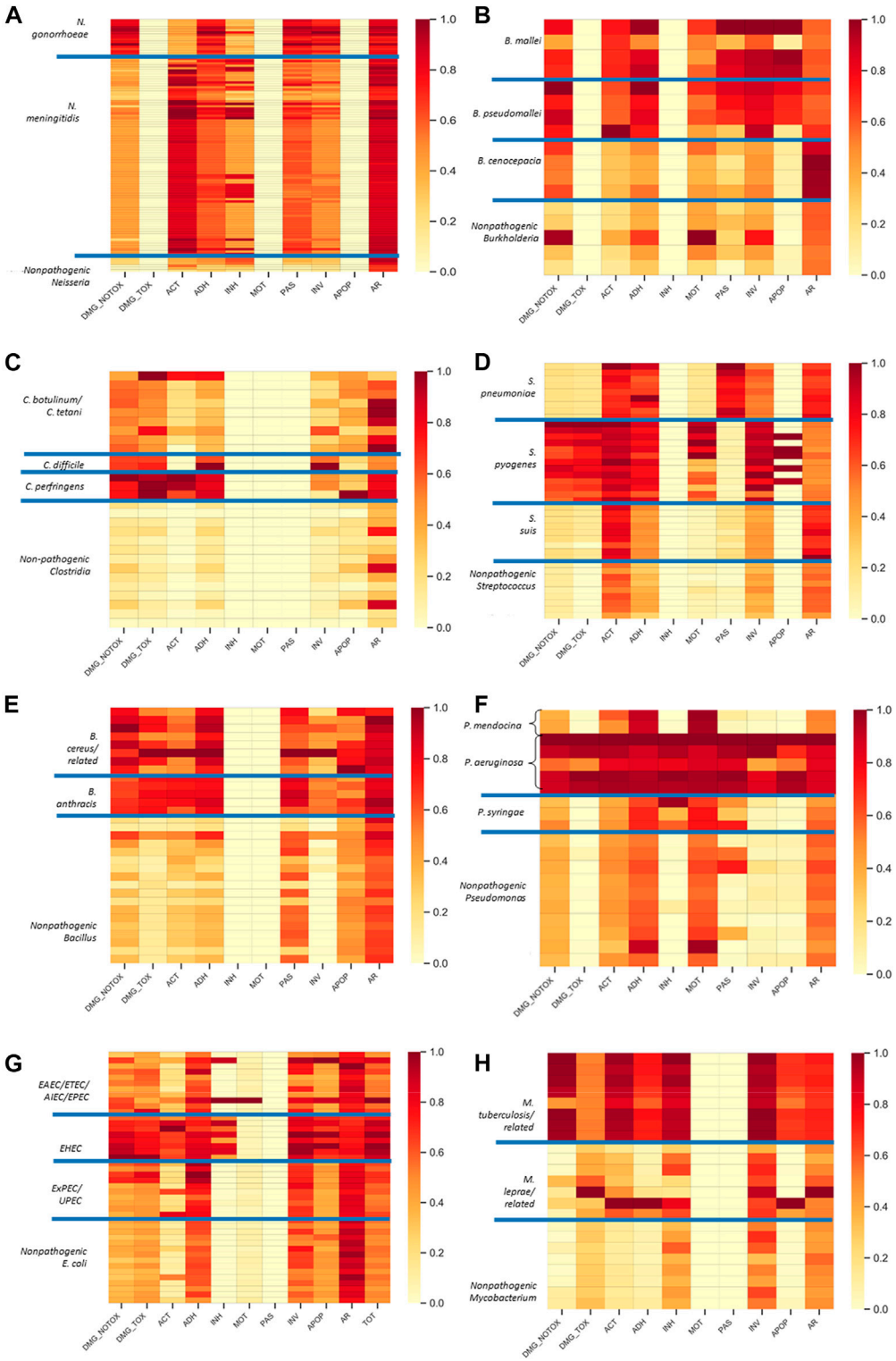

FIGURE 3
Pathogenic species are enriched in hazardous functional categories. Shown are the hazard fingerprints for Neisseria **(A)**, Burkholderia **(B)**, Clostridium **(C)**, Streptococcus **(D)**, Bacillus **(E)**, Pseudomonas **(F)**, *E. coli* **(G)**, and Mycobacterium **(H)**. The fingerprints are shown as rows in a heat plot with the values in each column representing the normalized fraction of CDSs within each functional category (as defined in Table 1). Only the relevant categories from Table 1 are included (i.e., those that provided alignments). The pathogenic subgroups within each organism group are defined in Table 3 and separated by the blue lines on the heat plot. Abbreviations: DMG_NOTOX, damage without toxin activity GO term, DMG_TOX, damage with toxin activity GO term; ACT, active host subversion; ADH, adherence; INH, inhibits host cell death; MOT, motility; PAS, passive host subversion; INV, invasion; APOP, host cell apoptosis; AR, antibiotic resistance; TOT, TOTAL (sum of all other categories).

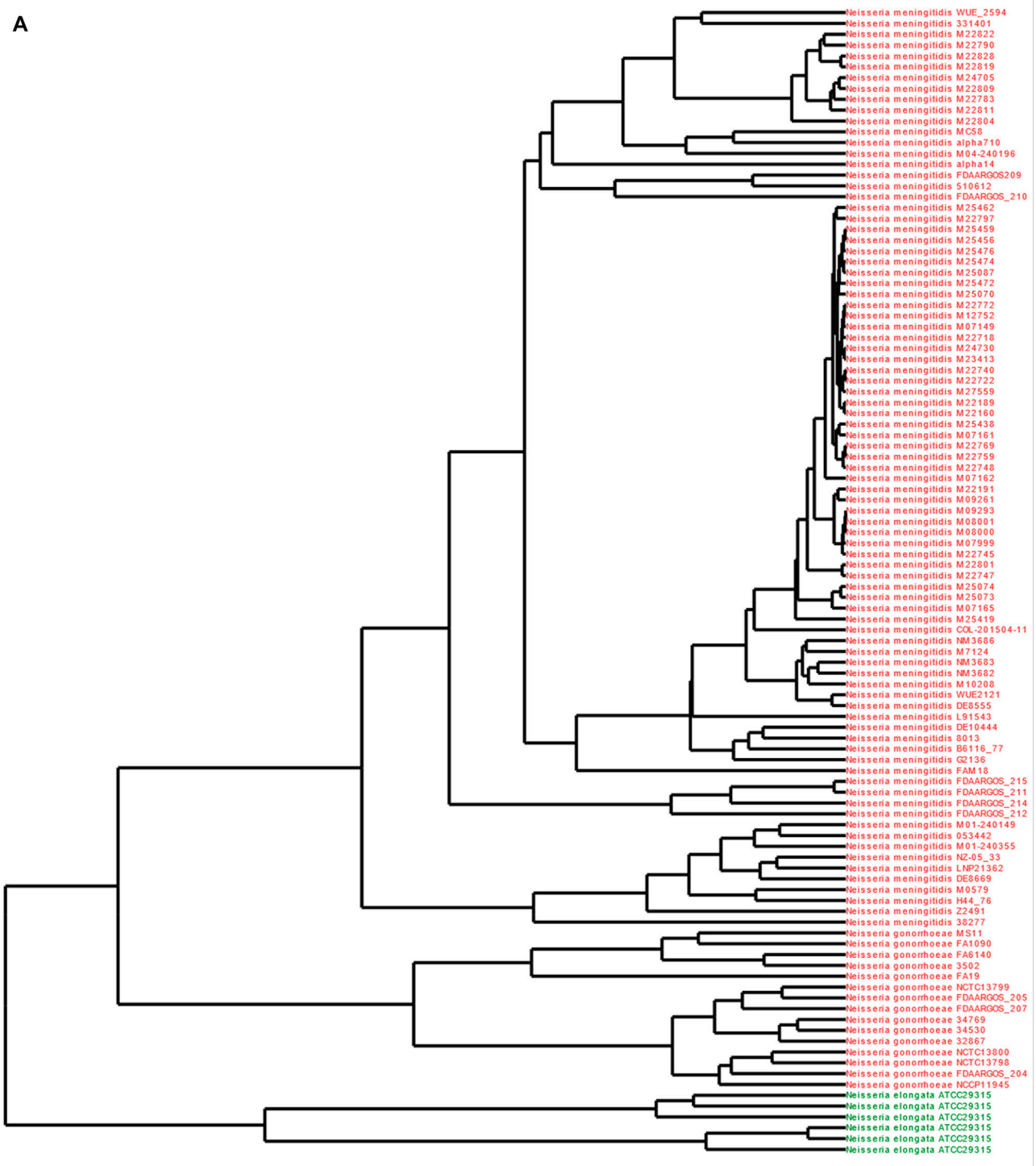

FIGURE 4
Hazardous functions separate pathogens from non-pathogens. Shown are the dendrograms for Neisseria **(A)**, Burkholderia **(B)**, Clostridium **(C)**, Streptococcus **(D)**, Bacillus **(E)**, Pseudomonas **(F)**, *E. coli* **(G)**, and Mycobacterium **(H)**, with pathogenic species colored in red and non-pathogens colored in green. An additional plot for *E. coli*, stratified by the groups shown in Figure 3 is shown in Supplementary Figure S1.

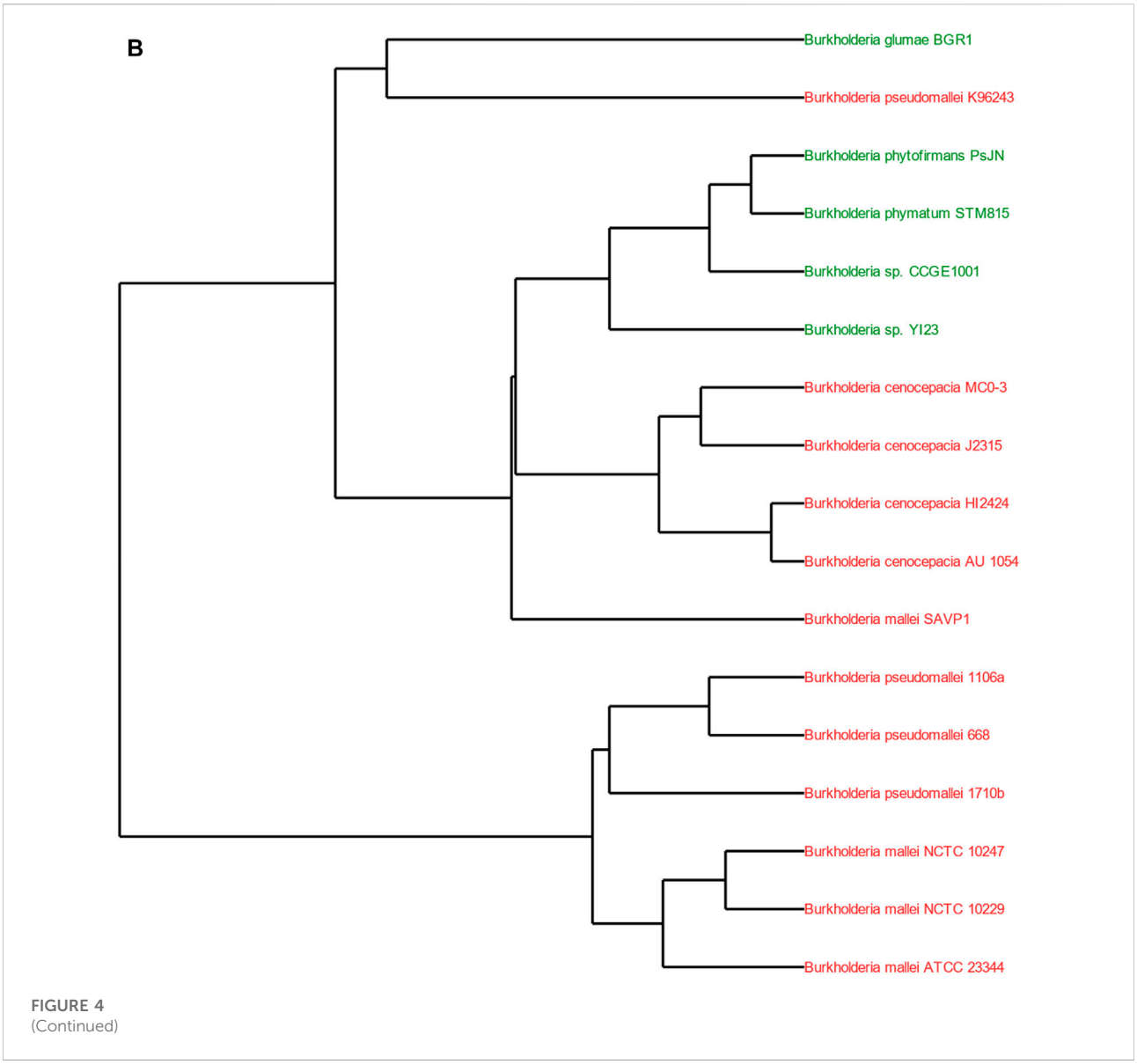

FIGURE 4
(Continued)

the heat plots using agglomerative hierarchical cluster analysis. These clusters were then visualized by plotting dendrograms, where known pathogenic groups were labeled in red, and non-pathogenic in green. For most organisms, hierarchical clustering based on the fingerprint data effectively distinguished between pathogenic and non-pathogenic strains Figure 4).

Overall, the plots demonstrate high levels of hazardous functions in pathogens relative to nonpathogens (Figure 3) and good separation between pathogen and non pathogens (Figure 4). More specifically for the fingerprints, there is good separation across most categories with the exception of antibiotic resistance, and the types of hazardous functions are consistent with literature reports as described below. For example, as shown in Figure 3A, both pathogenic Neisseria groups are enriched relative to the nonpathogen group in adherence, passive host subversion, and invasion functions. Further, the dendrogram demonstrates clear separation between pathogens and nonpathogens (Figure 4A). These findings are consistent with Lu et al., who demonstrated several genes unique to pathogenic Neisseria species that are involved in host immune evasion and adherence (Lu et al., 2019). *N. gonorrhoeae* further contains strains enriched in critical non-toxin damage functions, and *N. meningitidis* is enriched in active host subversion functions such Factor H binding protein (Supplementary Table S1).

Similarly, pathogenic Clostridium groups are clearly separated (Figure 4C), and pathogens are enriched in damage, adherence, and invasion functions relative to the nonpathogen group, with some strains being enriched in active host subversion and apoptosis (particularly the *C.*

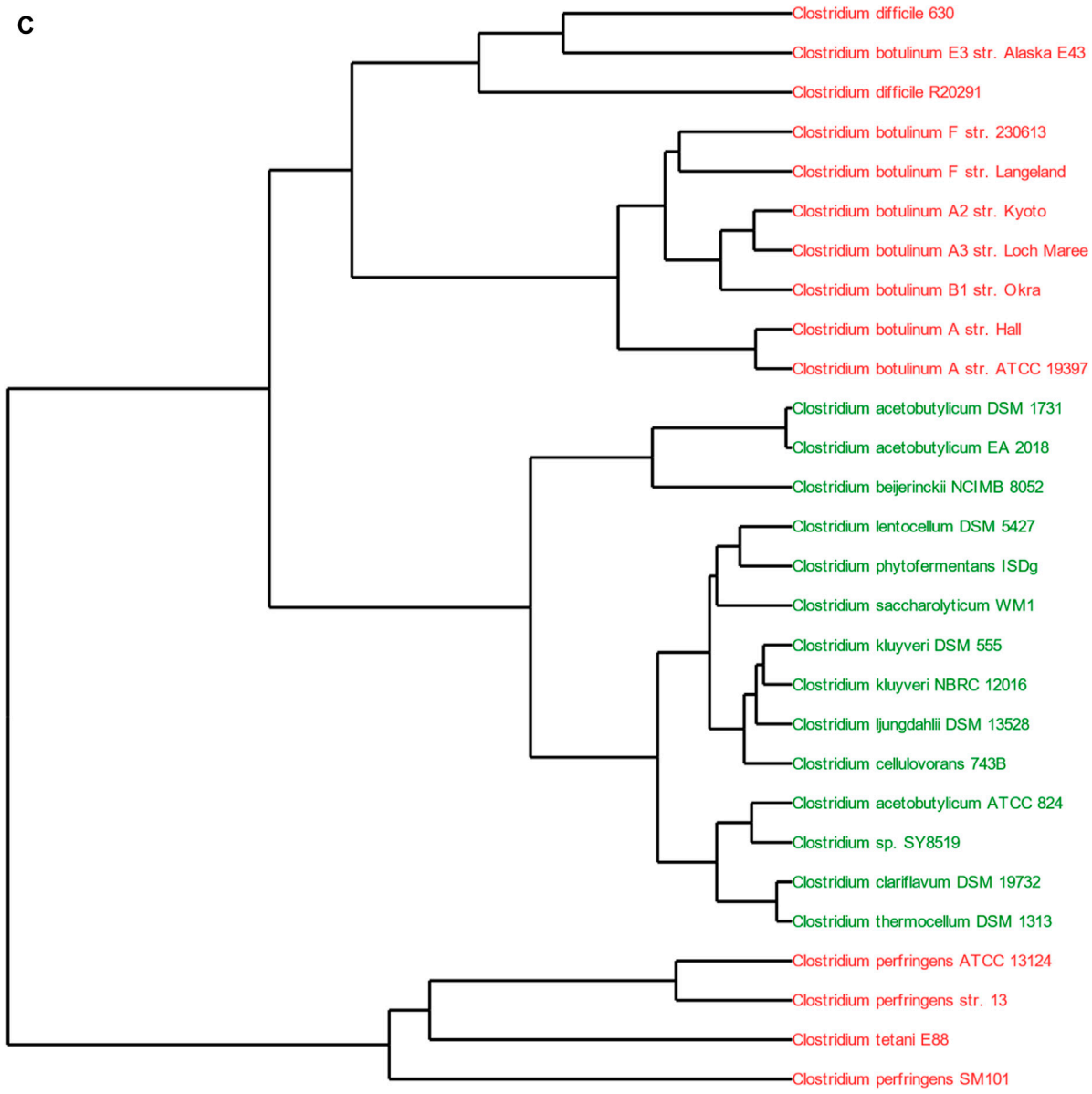

FIGURE 4
(Continued)

perfringens group) (Figure 3C). The most striking of these enriched categories for Clostridium are the damage categories, which is consistent with various Clostridium species producing damage-inducing factors such as toxins as their main hazardous functions, of which some can aggravate the immune response (Supplementary Table S1). For example, *C. botulinum* produces neurotoxins, *C. difficile* produces toxin A, toxin B, and binary toxin, and *C. perfringens* produces over 16 toxins (Awad et al., 2014; Rasool et al., 2017). Because the numbers of toxins produced by *C. perfringens* relative to the other two pathogenic groups is relatively higher compared to the other pathogenic groups, greater delineation between this pathogen group and the nonpathogenic Clostridium group is apparent due to the normalization process.

The Bacillus fingerprints (Figure 3E) demonstrates that Bacillus pathogens are enriched in functions related to damage, active host subversion and adherence relative to their nonpathogenic groups. The fingerprint plot also demonstrates that nonpathogenic Bacillus have antibiotic resistance functions,

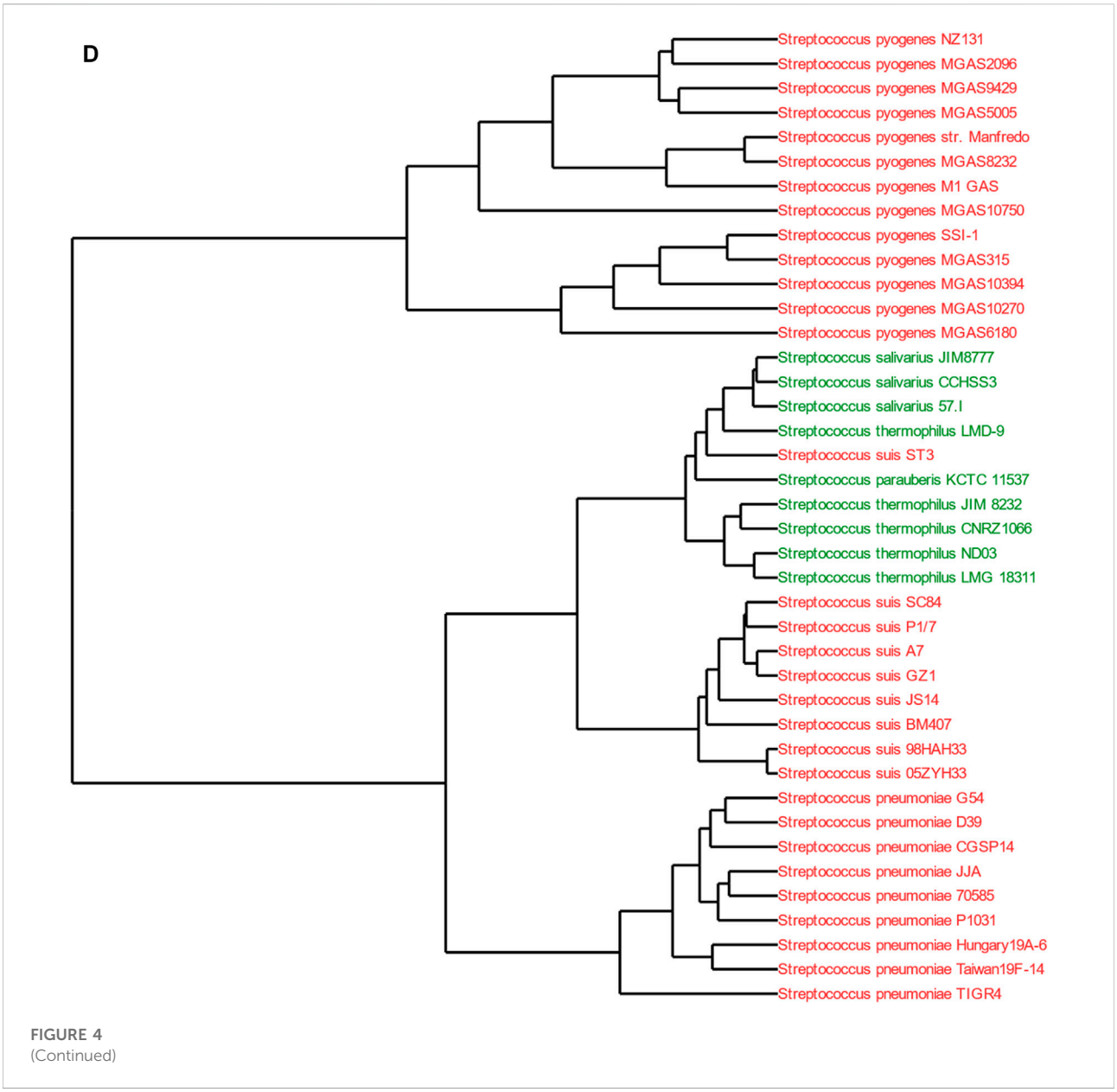

FIGURE 4
(Continued)

which supports other reports (Adimpong et al., 2012; Noor Uddin et al., 2015). For *B. anthracis*, the damage and active host subversion are most clearly delineated from the nonpathogen group, which is consistent with anthrax toxin—composed of protective antigen, edema factor and lethal factor (Supplementary Table S1)—being the major contributor to disease through destruction of host immune cells (Friebe et al., 2016; Visiello et al., 2016). Similarly, *B. cereus* contains factors that promote cell (including immune cell) damage, such as enterotoxins, hemolysins, emetic toxins, and phospholipases (Supplementary Table S1) (Visiello et al., 2016). Taken together, these functions allow separation of pathogens and non-pathogens (Figure 4E), with exception of one presumably non-pathogenic B. cereus strain Q1, an extremophilic strain known for microbial enhanced oil recovery due to production of biosurfactants (Xiong et al., 2009).

The plots also show good separation of some of the Streptococcus species from the nonpathogenic groups, particularly *S. pyogenes* (Figures 3, 4D). *S. pyogenes*—known as Group A Streptococcus clinically—has several factors enabling invasion, adherence, and motility within host cells, but perhaps the most important factors contributing to pathogenicity of *S. pyogenes* are the few proteins leading to direct damage (e.g., streptolysins O and S, and exotoxins A and C) and host evasion (e.g., IgG-degrading enzyme and Protein M) (Hamada et al., 2015). These critical functions are apparent

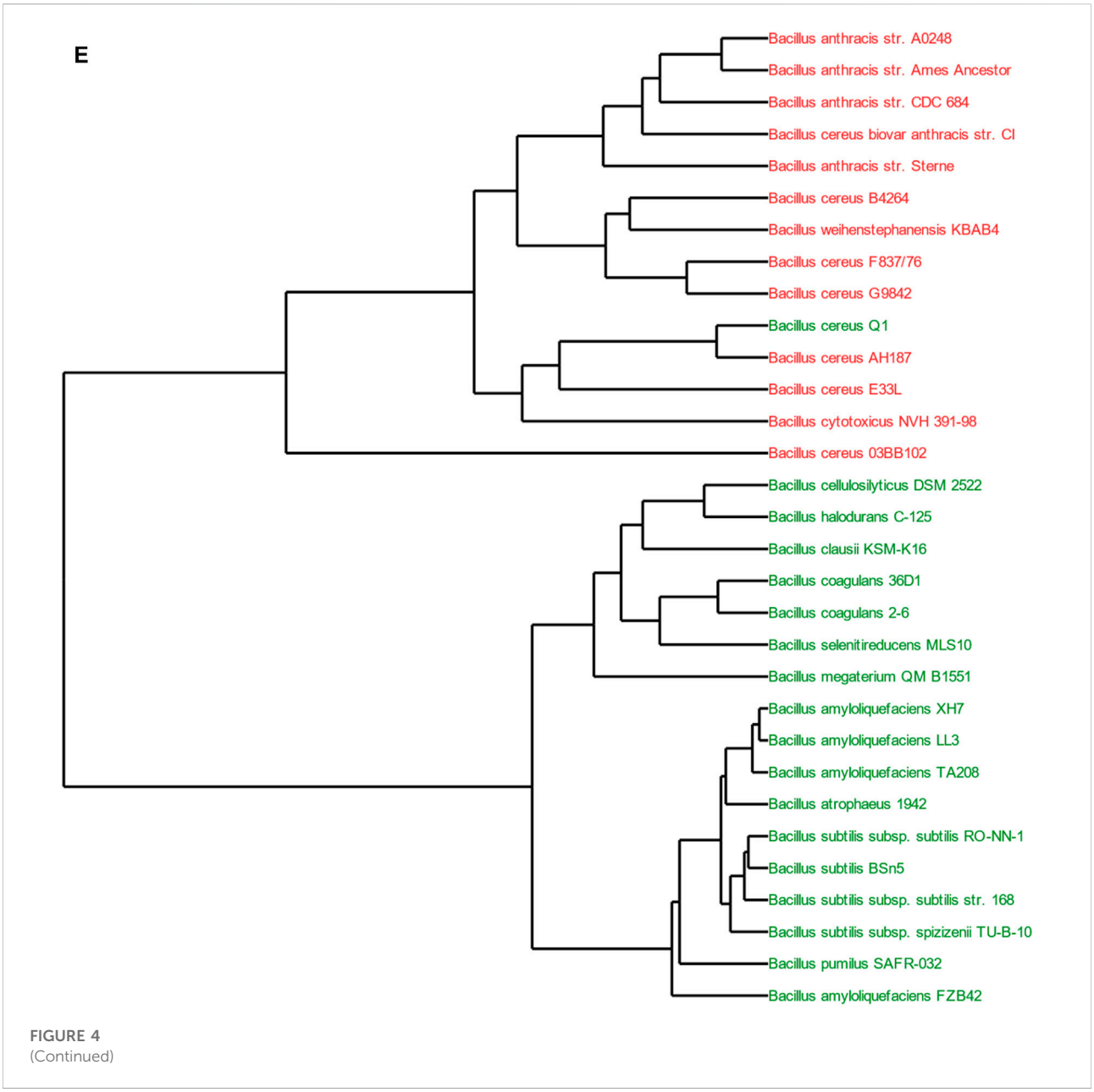

FIGURE 4
(Continued)

in the heat plot as well as Supplementary Table S1. Less defined separation is apparent between the nonpathogenic group and the *S. pneumoniae* or *S. suis* group with a few exceptions. For example, antibiotic resistance factors show some delineation from the nonpathogen and *S. pneumoniae* or *S. suis* groups, which is consistent with the emergence of antibiotic resistance strains in these species (Nuermberger and Bishai, 2004; Yongkiettrakul et al., 2019). Further, enzymes leading to *S. pneumoniae* cell wall decoration that enable immune system avoidance (Mitchell and Mitchell, 2010) likely contributes to this group being separated from the other groups within the passive immune subversion category. *S. pneumoniae* and *S. suis* also express critical damage factors, such as the PLY pore-forming toxin (Mitchell and Mitchell, 2010) and hemolysins (Haas and Grenier, 2018), respectively, which—while not very apparent in Figure 3 due to high levels of the damage functional category in *S. pyogenes*—are identified as critical factors in Supplementary Table S1. Taken together, these hazardous functions enable good separation of pathogens from non pathogens. One exception is the pathogenic strain *S. suis ST3*. According to this Hu et al., this strain is missing a large pathogenicity island (Hu et al., 2011), which is the likely cause of lack of separation.

Like Streptococcus pathogens, Mycobacterium pathogens, particularly tuberculosis-causing Mycobacteria, are separated

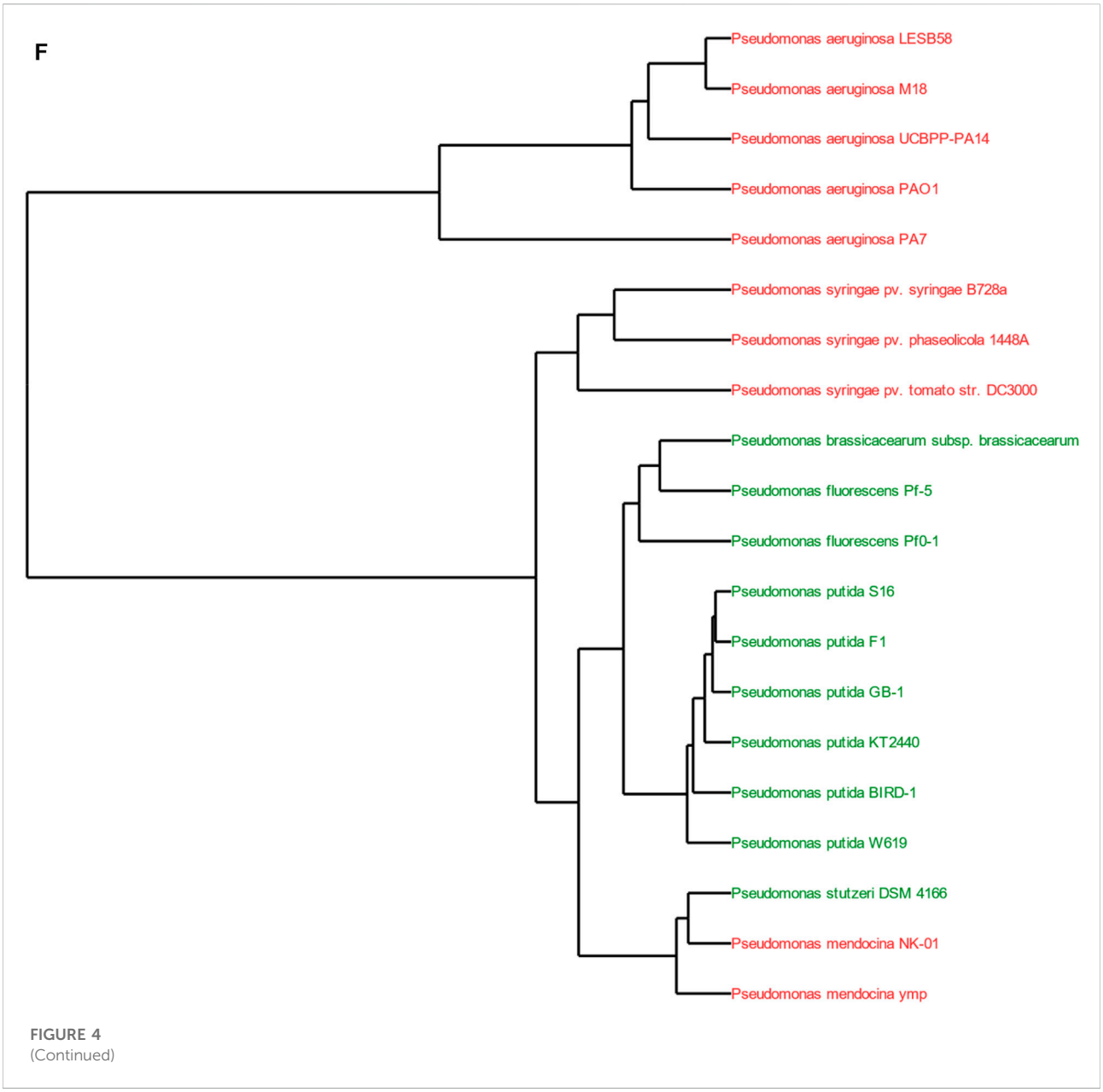

FIGURE 4
(Continued)

well within specific hazardous categories (Figure 3H) and separate well from non-pathogens (Figure 4H). One exception is *M. abscessus ATCC 19977*, a pathogen that clusters with non pathogens. This finding is actually consistent with another report, which demonstrated that this strain clusters with other non-pathogens based on whole proteome analysis (Zakham et al., 2012). In general, we found that *M. tuberculosis* strains are enriched in active host subversion, adherence, and apoptosis categories relative to the nonpathogen group, which is consistent with the fact that *M. tuberculosis* virulence largely depends on the organism's ability to infect host cells and evade the host immune response (Forrellad et al., 2013). The plot additionally shows that damage factors contribute to differences compared to the nonpathogen group, which supports the fact that *M. tuberculosis* requires damage factors such as adenylate cyclase (Supplementary Table S1) for virulence (Agarwal et al., 2009). In contrast to *M. tuberculosis*, less separation is apparent for the *M. leprae* and related group. This observation is likely because only 24 of the 339 Mycobacterium hazardous functions contained in our database are from the *M. leprae* and related group, and the CDSs from this group may not have enough homology to hazardous functions from *M. tuberculosis* strains to be relevant in our analysis.

Similar to the Mycobacterium analyses, some hazardous categories are emphasized for *E. coli*, although our analysis

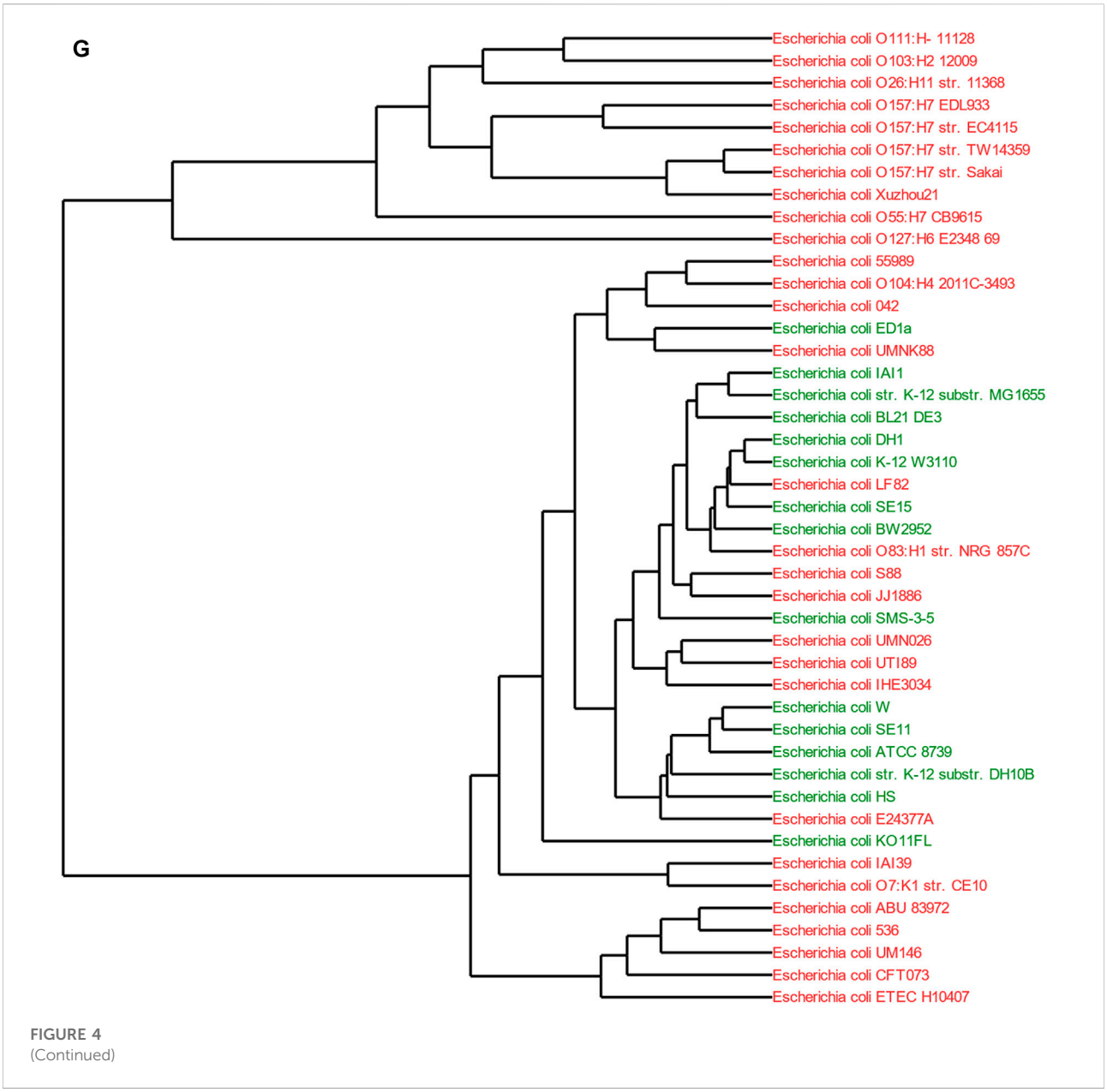

FIGURE 4
(Continued)

was not able to clearly separate all pathogenic groups (Note: Figure 4G colors and labels the dendrogram based on pathogenic and non-pathogenic strains, whereas Supplementary Figure S1 colors by pathogenic and non-pathogenic group). Since infections caused from intestinal pathogenic *E. coli* (IPEC) are distinct from infections caused extraintestinal pathogenic *E. coli* (ExPEC, including uropathogenic *E. coli*) (Kohler and Dobrindt, 2011), we separated with *E. coli* pathogenic strains into IPEC strains—including a group of enterohaemorrhagic *E. coli* (EHEC) and non-EHEC strains (EAEC/ETEC/AIEC/EPEC)—and ExPEC strains. While EHEC strains are clearly separated (Supplementary Figure S1), ExPEC strains could not be separated as well, likely because these strains can belong to the normal (nonpathogenic) gut flora and share large portions of their genome with nonpathogenic strains (Kohler and Dobrindt, 2011). In contrast to the ExPEC strains, the IPEC strains—particularly the EHEC strains—show greater relative abundance of damage functions (Figure 3E). This observation supports that fact that functions that contribute to host cell damage are critical to IPEC pathogenesis, such as enterotoxins and shigatoxins (within ETEC and EHEC strains, respectively) as well as functions leading to attaching and effacing lesions (Welch et al., 2002; Kaur et al., 2010; Nguyen and Sperandio, 2012). The EHEC group is also further differentiated from the other IPEC strains within the active host subversion and inhibits host cell death categories, which is a hallmark of EHEC strains (Ho et al.,

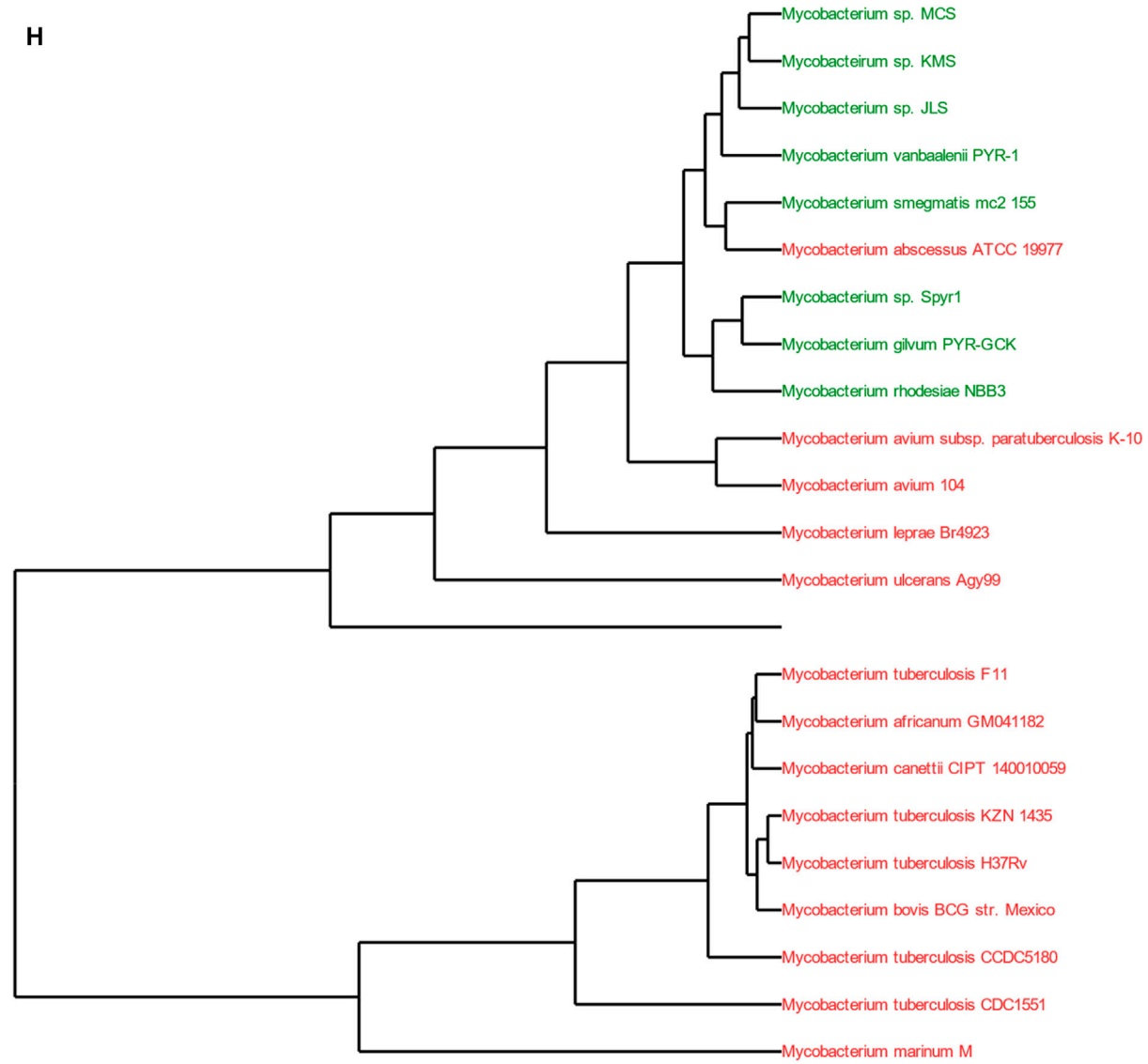

FIGURE 4
(Continued)

2013). IPEC strains also elicit aggressive adherence functions to enable pathogenicity (Kaur et al., 2010), but our methods did not enable clear emphasis of this category in pathogenic strains compared to nonpathogenic stains, likely due to the ubiquitous nature of adherence functions.

For Burkholderia, our analysis enables good separation, with the exception of B. pseudomallei K96243, a pathogen that clusters with non-pathogens (Figure 4B). Previous analysis of the genome of this strain noted high similarity to Ralstonia solanacearum, a plant pathogen (Holden et al., 2004), which is consistent with this strain clustering with B. glumae and B. phytofirmans (plant colonizers) in our analysis. *B. mallei* and *B. pseudomallei* are intracellular pathogens that use numerous virulence factors that enable host cell survival, such as invasion and immune evasion factors (Galyov et al., 2010; Memisevic et al., 2014), which is apparent in Figure 3B. These organisms also contain key factors such as BimA, hemagglutinin, PilA, which are involved in invasion, damage, and adherence, respectively (Sarovich et al., 2014) that enable emphasis of these categories in the plot. In contrast to *B. mallei* and *B. pseudomallei*, the only enriched functions for *B. cenocepacia* are antibiotic resistance and non-toxin damage functions, but this may be an indication of lack of coverage in our database (only 2 of the 141 hazardous Burkholderia functions are from *B. cenocepacia*). However, this finding is consistent with the fact that *B. cenocepacia* clinical strains isolated from cystic fibrosis patients

10.3389/fbioe.2022.979497

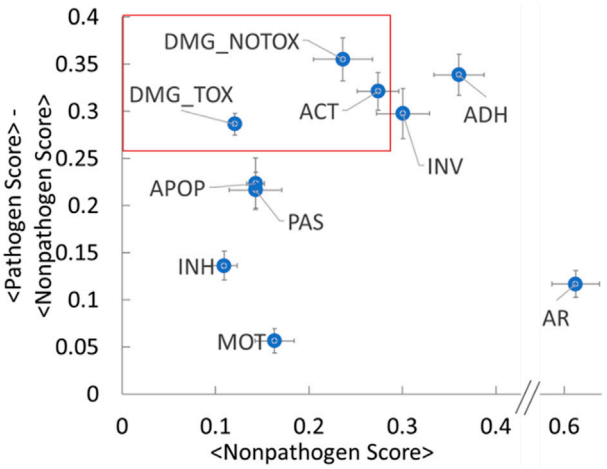

FIGURE 5
Pathogen and Nonpathogen Fingerprint Scores Reveal Stratification Among Functional Hazard Categories. The plot shows a difference in average scores between the pathogens and nonpathogens (*y*-axis) as a function of the average nonpathogen score (*x*-axis) for the functional categories from the heat plots in Figure 3. The error bars represent the standard error. The abbreviations are the same as Figure 3.

can be resistant to antibiotics and contain several lipases and proteases to illicit tissue damage (Mahenthiralingam and Vandamme, 2005). Noticeably, *B. glumae* (third row from the bottom in Figure 3B demonstrates some pathogenic signatures, which is consistent with research demonstrating that this species can be a rice pathogen (Pedraza et al., 2018). This species was originally considered a nonpathogen based on the dataset published by Cosentino et al. (Cosentino et al., 2013), suggesting that our methods may enable identification of misannotated organisms.

Finally, some separation is also apparent for Pseudomonas species, but the patterns are not as consistent across strains as the other pathogens (Figures 3, 4F). Pseudomonas species pathogenic to humans (*P. aeruginosa* and *P. mendocina*) have a wide variety of virulence factors (Goldberg, 2010), but the patterns are different between the two species, and these two groups are completely separated in the dendrograms (Figure 4). For example, both *P. aeruginosa* and *P. medocina* have several proteins contributing to adherence and motility (Supplementary Table S1), but these types of functions can occur in nonpathogenic species as well. In contrast, invasion factors, host cell subversion factors, host cell apoptosis, and damage factors are relatively unique to *P. aeruginosa* strains (Figure 3 and Supplementary Table S1), which is consistent with experimental evidence (Shaver and Hauser, 2004; Dulon et al., 2005; Casilag et al., 2016; Basso et al., 2017; Reboud et al., 2017). Antibiotic-resistance functions are higher in *P. aeruginosa* pathogenic strains as well, which is consistent with the clinical prevalence of antibiotic resistant strains (Jacoby and Munoz-Price, 2005). For plant pathogens, our methods result in some separation of *P. syringae*—a plant pathogen—from nonpathogenic Pseudomonas species overall (Figure 4), and within the inhibits host cell death functional category (Figure 3). These observations may be driven by the fact that only 2 of the 175 Pseudomonas hazardous functions contained in our database are from *P. syringe*.

Toward application of the methodology and resulting functional hazard database

The fingerprint analysis presented in the previous section demonstrates that categorizing hazardous functions allows the importance of the gross functionalities (i.e., the functional metadata categories in Table 1) to differentiate nonpathogenic groups from pathogenic groups for both gram-negative and gram-positive bacteria. As further demonstration of our methodology and database with an eye toward the utility of our method for biosafety assessments, we sought to determine the relative hazard level of each functional category. Logic suggests that two parameters play a large role in such a relative ranking: 1) the magnitude of the category's increase in relative abundance compared to nonpathogens and 2) the relative abundance of the category in nonpathogens. As a simple measure of these parameters, we leverage the data used to generate the heat plots to calculate an average score for each of the functional categories for the nonpathogen and pathogen groups. Figure 5 shows a plot of the difference in average scores between the pathogens and nonpathogens as a function of the average nonpathogen score. The points on the upper left quadrant of this graph thus represent highly hazardous categories that 1) have a relatively large difference between the pathogen and nonpathogen scores and 2) have a low background signature (i.e., low nonpathogen score). For example, these results suggest that the damage (with and without toxin activity) and active host subversion categories have relatively high pathogen-nonpathogen difference scores (e.g., >0.25) with low nonpathogen scores (e.g., <0.3) (red box in Figure 5). Such an analysis demonstrates a potential ranking system for "sequences of concern," and may enable a foundation for a risk-based approach for biohazard assessments for designed organisms. As mentioned above, more hazardous functions that do direct damage to a cell or those involved in avoiding the host immune system rank more highly than less hazardous functions such as adherence and motility. Thus, the damage and active host subversion categories may present a higher hazard relative to other categories for biohazard analysis, for example. Generalizing this approach across all functional categories and all organism types may provide an objective foundation for biohazard analysis of novel organisms.

Discussion

While the methodology and database presented here has two immediate uses—1) biosecurity screening assessments of synthetic

genes and 2) partial biosafety assessments for bacterial genomes—future work should build upon this foundation to provide comprehensive biosecurity and biosafety assessments for the synthetic biology community. We envision a future in which any novel biodesign can be assessed through a function-based paradigm that requires only genomic sequences. This paradigm is in contrast to current biosafety assessments that rely on phenotypic information from well characterized organisms to classify organisms into Biosafety Levels, for example, which provides researchers with an understanding of the level of pathogenicity, transmissibility, and other characteristics of the organism (U.S. Department of Health and Human Services, 2014). However, as the genomes of new biodesigns begin to deviate further and further from these well characterized organisms, biosafety levels become less and less clear, thus necessitating in silico genome characterization methods. Where traditional biosafety assessments are limited to known pathogens with no or minimal bioengineered parts, with future development, our framework may enable assessment of seemingly limitless potential for biodesigned organisms. In this discussion, we elaborate on the issues with the current paradigm, how our approach begins to shift the paradigm, and the future work needed to provide a complete paradigm shift.

Progress in bioengineering, synthetic biology, and computational science is enabling artificial creation (*de novo* genetic synthesis) of whole organisms, including viruses (Blight et al., 2000; Cello et al., 2002; Smith et al., 2003; Oldfield et al., 2017; Noyce et al., 2018) and bacteria (Gibson et al., 2010; Hutchison et al., 2016), as well as recombinant production, viral reverse genetics, rational design, design from standardized DNA components (e.g., Biobricks), and/or modular protein assembly (e.g., SpyTag or SpyCatcher (Khairil Anuar et al., 2019)). Such technologies have led to exponential growth of publications based on synthetic biology since 2000, and larger throughput per synthetic biology lab (Raimbault et al., 2016). Further yet, DNA synthesis is becoming more distributed, for instance, with the availability of DNA printers such as the BioXp system from Codex DNA. As breakthroughs are made to realize the promise of synthetic biology, the creation of novel sequences may expand even more, and such growth is difficult to monitor. Although the numbers of new natural strains being discovered is accelerating fairly linearly (Suzek et al., 2015; RefSeq, 2019), the production of bioengineered strains may be growing exponentially, as many of these sequences are not publicly available. This rapid progress in bioengineering has created a gap in current biosafety practices that requires a framework to understand the potential hazards posed by functional building blocks. We have provided empirical data that demonstrates a function-centric paradigm for identifying and classifying hazardous biological parts. The functional classification of sequences is based on coarse hazardous functions encoded by organisms, such as functions contributing to pathogenicity, toxin and drug production, and immune regulation.

The methodology demonstrated here can immediately be used for partial biosafety assessments for bacterial genomes for classification of pathogens and non-pathogens using functional hazard fingerprints. Future iterations of the method should involve testing both previously characterized organisms and novel organisms (i.e., those not contained in the database and/or novel biodesigns with known phenotypes) in order to characterize a variety of biosafety-related characteristics (not just pathogenic/nonpathogenic) from various domains of life beyond bacteria (e.g., viruses and fungi). As we demonstrated in Figure 4, hierarchical clustering achieves a high level of separation between pathogen and nonpathogen organism group members using a simple alignment with default parameters against our curated database. This approach is in contrast with more complicated, manual annotation and phylogenetic analysis that require time-consuming, expert interpretation. Even outlier pathogens that cluster with nonpathogens like *S. suis ST3* have characteristics that explain why they do not cluster with other pathogens; for example, as noted *S. suis ST3* clusters with nonpathogenic organisms but is missing a pathogenicity island, which likely contains several hazardous functions. Similarly, outlier nonpathogens that cluster with pathogens such as *B. cereus Q1* can be explained as well. The genome for this organism contains genes encoding for enterotoxins (NCBI accessions ACM12308, ACM12309, and ACM12310) involved in damage and adherence, lipid transferases involved in passive and active immune subversion (accessions ACM11963 and ACM12924) and antibiotic resistance (accessions ACM12845 ACM12455). Thus, if used for assessments of pathogenicity, false negatives (due to lack of hazardous functions and/or presence of previously uncharacterized functions) and false positives (due to the presence of hazardous pseudogenes and/or non-hazardous sequences with high homology to hazardous functions) could occur depending on the thresholds used for classification. However, the success of this approach demonstrates the native utility of the hazardous function database and that further refinements in fingerprinting approach are both attainable and could be an effective diagnostic approach to classifying unknown organisms.

As documented in Table 3, some pathogens have higher coverage in our database than others, and thus comparison across pathogenic groups should be interepreted appropriately. Differences between a pathogen and nonpathogen in one organism being less pronounced relative to another organism group could be due to large functional differences, but it could also be due to lack of database coverage. For example, the fact that *M. tuberculosis* pathogens have higher numbers of hazardous functions compared to *N. gonorrhoeae* does not mean necessarily that *N. gonorrhoeae* is relatively more pathogenic compared to its nonpathogenic conterparts than *M. tuberculosis*; this result may be driven by the larger coverage of Mycobacterium sequences within our database. Determination if our approach can be used

to elucidate levels of pathogenicity based on a collection of hazardous functions warrants further exploration. Such an application may have utility beyond biosafety assessments, such as emerging and recurrent disease identification. As recently stated by others, new approaches are needed to address emerging diseases (Reperant and Osterhaus, 2017), particularly as surveillance and diagnostics improve across the globe. We propose that a function-based paradigm provides a foundation to meet this need, and such approaches have already shown success. In this study, we leveraged data from Cosentino et al., who developed methods to classify bacterial pathogens from nonpathogenic bacteria based on protein families (Cosentino et al., 2013), which have a direct link to function (Pearson, 2013). Beyond bacteria, others have shown that sequence differences leading to functional differences are critical determinants of pathogenicity for viruses and fungi such as influenza virus (Ebrahimi et al., 2014; Straus and Whittaker, 2017), African Swine Fever Virus (Chapman et al., 2008), Zika virus (Shah et al., 2018), Colletotrichum spp. (Vieira et al., 2019), and Geosmithia spp. (Schuelke et al., 2017). Thus, development and generalization of models may aid in the shift from organism to function-based classifications for all types of infectious disease. For example, a logical extension of the study presented here would be to determine if similar results can be obtained if we leveraged our entire database (not just specific subsets of hazardous functions from selected bacteria), such that a prior knowledge of the organism in question is not needed.

In addition to the immediate use of our methods for predicting pathogenicity of bacteria, the method and database also has immediate use for screening individual gene sequences. The example application of our methodology and database to stratify sequences are promising, but the results suggest more granular functional categories may be needed to enable use for more pointed biosafety assessments. Granular metadata for protein sequences are available from several databases that are cross-referenced within the UniProt Knowledge Database (UniProt, 2019b), such as Gene Ontology (GO) terms (Ashburner et al., 2000; The, 2019), Interpro terms (Mitchell et al., 2019) and sequence features (e.g., motifs, regions, mutation impact, etc.). GO terms provide a graphical representation of molecular functions, biological processes, and cellular components of gene products and their relations among each other (Ashburner et al., 2000; The, 2019). We leveraged the "toxin activity" GO term within our framework, but further use of GO terms may enable better stratification of hazardous sequences.

Our results may also improve if host information is considered. Recent efforts, such as ViralZone (ViralZone, 2019) and the proposed PathGO (IARPA. Broad Agency Announcement, 2016) are providing better GO terms for host-pathogen interactions that may prove valuable for function-based hazard classification. Casadevall proposed a damage response framework (Casadevall and Pirofski, 2003) that is founded on the simple principle that microbial pathogenesis is "the outcome of an interaction between a host and a microorganism" measured by damage to the host. Current knowledge suggests pathogens interact with the host in a variety of ways, including mimicking host activities, leading to a lack of host cellular control (Knodler et al., 2001; Stebbins and Galan, 2001; Smatti et al., 2019), but documentation of these data in a machine-readable format is sparse. Two potentially useful sources of information that are cross-reference in UniProt are IntACT (Hermjakob et al., 2004), which provides protein-protein interaction data, and Reactome (Reactome, 2019), which provides functional metadata associated with biological pathways. An initial analysis of our hazardous functions suggests that <2% of protein accessions in our database have at least one interactor in IntACT database, and 58% of the interacting proteins are human proteins. These human proteins represent 3% of the total reactome metadata. In addition to IntACT, specific (host-pathogen) protein-protein interaction information is available from Biogrid (Oughtred et al., 2019), String (von Mering et al., 2005) and other databases, but information is sparse. However, as high-throughput experimentation becomes more commonplace, information contained in these databases can be leveraged for hazard analyses. Specifically, further expansion of these databases for hazardous sequences may be needed for impactful analysis and utility into a function-based biosafety assessment.

In addition to hazards that may impact hosts such as humans, livestock, and crops, other living hosts and non-living "hosts" of economic importance should be considered as well for other pointed biosafety assessments. For example, when considering safety assessments for novel bio-based fertilizers and/or biopesticides, hazards with economic impact potential beyond those that effect crops and livestock may need to be considered. For example, of the world's ~250,000 flower and seed-producing plant species, between 78% and 94% require pollinators for fertilization ("FAOSTAT" Food and Agricultural Organization www.fao.org), with bees accounting for pollination of approximately 30% of the world's food supply (Klein et al., 2007). Bee colonies can collapse from fungal, bacterial and viral outbreaks, such as those caused by the picornavirus-like deformed wing virus (DWV) and the ectoparasitic mite Varroa destructor (Tehel et al., 2019). Similarly, functions that could negatively impact non-eukaryotic or non-living "hosts" of economic importance should also be considered for tailored safety assessments. For example, under the current paradigm of biosecurity, biodesigns have been created that could potentially impact biomanufacturing supply chains (Abdulamir et al., 2014), control of pharmaceuticals (Galanie et al., 2015; Nakagawa et al., 2016), and crude oil supplies (Xu et al., 2018). Thus, as bioengineering rapidly progresses, safety practices need to keep pace to not only protect humans, livestock, and crops, but also protect infrastructure of critical economic impact.

Expansion of sequences and metadata may thus improve upon our foundation for biosafety practices of the bioengineering-centric future. Our methods and database reported here provide an understanding of the hazard posed

by "parts" of the organism, such that a foundation can be set to understand the hazard of the "whole." For example, *P. aeruginosa* has numerous hazardous functional parts including those contributing to adherence (type 4 pili and flagella for interacting host cells), invasion (T3SS), host cell subversion (biofilm formation, stimulation of proinflammatory response, and disabling of protease activity receptor-2), host cell apoptosis (exotoxin A stimulation of programmed cell death), damage (and cytotoxic effector proteins) and antibiotic resistance (beta-lactamases) (Shaver and Hauser, 2004; Dulon et al., 2005; Jacoby and Munoz-Price, 2005; Casilag et al., 2016; Basso et al., 2017; Reboud et al., 2017; Shen et al., 2017). While many of the hazardous functions of *P. aeruginosa* are known, a biodesign created with similar hazardous functions may not be identified under the current organism-centric paradigm. We must now build upon our methods developed using the engineering-like principle of pathogens being an organized assembly of functional hazards. Using this paradigm, we can then classify groups of sequences that compose a novel pathogen, thus enabling generalized function-based biosafety assessments for novel organism-level biodesigns for all types of applications.

Methods

Hazardous function database

Hazardous functions were identified from publicly available literature and databases (e.g., Supplementary Table S2) as those that have a function that impacts human and non-human hosts of high economic value as described in the Results section. We defined a hazardous function as a set of one or more protein sequences and associated manually curated metadata (Table 1). Each hazardous function can contain one or more functional categories. A hazardous function is only included in the database if its sequence encodes for a verified function based on experimental data from the literature or (in cases such as some select agent viruses where experimental data do not exist) based on homology to a sequence with verified function. Protein sequences were retrieved from UniProt when available or manually entered based on literature documentation. Functional metadata categories were developed based on panel discussions of high-level hazardous functions used by pathogens and organisms producing toxins, drugs, and bioregulators. For hazardous functions in the "damage," category, the toxin activity gene ontology term (GO:0090729) was used to distinguish toxins from non-toxins. Further, for sequences involved in the biosynthesis of small molecule toxins or drugs, hazardous functions were annotated with the step removed from the final product (e.g., last step, second-to-last step) based on pathway information as described in the literature and/or on Metacyc (Caspi et al., 2018).

Identification of hazardous coding sequences from bacteria

To validate the above methods and resultant database, we compared pathogenic and non-pathogenic strains against our functional hazard database. For this exercise, we compiled coding sequences (CDSs) from human and animal pathogenic and nonpathogenic strains based on the references outlined in Table 3. For each identified reference, pathogenic and nonpathogenic strains were reviewed; if a nonpathogenic strain was revealed as a pathogenic strain to a host of interest (or vice versa) based on other literature sources (e.g., a source published after the primary reference), it was removed from the analysis. Further, if an organism has known plasmids with sequences not deposited in NCBI, it was removed from the analysis. Pathogenic species or strains from each organism group were further stratified into subgroups based on species groups or disease-causing metadata (Table 3, column 4) for comparative purposes. CDSs, including those from chromosomal accessions and associated plasmid accessions were downloaded using NCBI's Batch Entrez online tool (NCBI, 2019). Plasmids were included since genetic determinants of bacterial virulence are often carried on mobile elements such as transposons and plasmids (Zaluga et al., 2014). Each strain's CDSs were defined by those contained within all chromosomes and plasmids associated with that strain. For each organism group, CDSs were aligned against a database of hazardous functions from its same genus using the Local Aligner for Massive Biological DatA (Lambda) (Hauswedell et al., 2014) version 2–1.9.5 using default settings. The alignment score, A, was defined as

$$A = Percent\ Identity \times Percent\ Hazardous\ Sequence\ Coverage \quad (1)$$

As discussed, we define the minimal alignment score for a CDS to be a hazardous function as 40% based on the thresholds used to define UniRef50 clusters. We then determined the fraction of hazardous CDSs (total number of CDSs in each strain normalized by the strain's total number of CDSs) and averaged the results of each strain within each pathogen and nonpathogen group.

Hazardous function fingerprinting

To determine a hazard function fingerprint for each strain, the alignment scores, A, for each CDS (to the genus-specific hazardous function database) were summed for each functional category then normalized to the maximal value across all pathogen and nonpathogen groups within that functional category. If a strain did not have a CDS with an alignment to the hazardous function, A was set to zero. Since each hazardous function can contain one or more functional categories, we defined the fingerprints as follows.

For each CDS set of alignment results (i.e., one CDS to one or more hazardous functions), the maximal A for each functional category (Table 1) was tabulated. For example, suppose CDS_i aligns to $Hazardous\ Function\ Sequence_1$ and $Hazardous\ Function\ Sequence_2$ with an A of 1.0 and 0.8, respectively. If $Hazardous\ Function\ Sequence_1$ has adherence metadata and $Hazardous\ Function\ Sequence_2$ has both adherence and invasion metadata, the fingerprint score contribution for CDS_i would be 1.0 for adherence and 0.8 for invasion. Maximum A scores for each functional category for each strain were then summed across each strain's CDSs. The final fingerprint score for each strain was defined as the cumulative A within each category normalized by the strain's total number of CDSs then normalized by the maximal value across all pathogen and nonpathogen strains within that functional category.

Hierarchical clustering analysis was performed in R using the function hclust, with UPGMA as the method for agglomerative clustering. Dendrograms were plotted using the R libraries ggdendro and ggplot2.

Authors note

The authors have carefully reviewed and discussed the concepts in this manuscript for dual use concerns both internally as well as with members of the US Government, the International Gene Synthesis Consortium (IGSC), and Engineering Research Council (EBRC). While we understand the risks, the prevailing opinion is that the methodologies presented here in themselves do not provide a roadmap for creation of harmful organisms, nor do they enable circumvention of screening. In fact, this manuscript provides the scientific community a potential framework for screening, which should help improve biosecurity through improved screening practices. Further, the authors purposefully did not publicize our database to further alleviate such concerns.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

CB contributed to the conceptualization of the paper and writing. BG contributed to curation, writing, and analysis. OT contributed to conceptualization and review. CM contributed to analysis and review. CH, DH, ZS, and LH contributed to data curation.

Funding

This work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), *via* contract number W911NF-17-C-0052. 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, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Acknowledgments

A special thanks to Gene Godbold, Sara Nitcher, Rachel Spurbeck, Meg Howard, Morris Makobongo, Nikolas Kanel, David Eaton, and Brett Fowle for their contributions to populating information into the biological functions hazard database and developing software for automated hazard level predictions based on protein metadata.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RM declared a shared research group with the author CB to the handling editor.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2022.979497/full#supplementary-material

Hazardous Functions Partially Separate *E. coli* Pathogen Groups Shown are the dendrograms for *E. coli* grouped by type of *E. coli*. Pathogenic species colored as follows: EHEC (red), ExPec/UPEC (purple), EAEC/ETEC/AIEC/EPEC (orange). Non-pathogenic species are colored as follows: commensal (green and teal) and yellow (lab strains).

References

Abdulamir, A. S., Jassim, S. A., and Abu Bakar, F. (2014). Novel approach of using a cocktail of designed bacteriophages against gut pathogenic *E. coli* for bacterial load biocontrol. *Ann. Clin. Microbiol. Antimicrob.* 13, 39. doi:10.1186/s12941-014-0039-z

Adimpong, D. B., Sorensen, K. I., Thorsen, L., Stuer-Lauridsen, B., Abdelgadir, W. S., Nielsen, D. S., et al. (2012). Antimicrobial susceptibility of Bacillus strains isolated from primary starters for African traditional bread production and characterization of the bacitracin operon and bacitracin biosynthesis. *Appl. Environ. Microbiol.* 78, 7903–7914. doi:10.1128/aem.00730-12

Agarwal, N., Lamichhane, G., Gupta, R., Nolan, S., and Bishai, W. R. (2009). Cyclic AMP intoxication of macrophages by a *Mycobacterium tuberculosis* adenylate cyclase. *Nature* 460, 98–102. doi:10.1038/nature08123

Ahr, B., Robert-Hebmann, V., Devaux, C., and Biard-Piechaczyk, M. (2004). Apoptosis of uninfected cells induced by HIV envelope glycoproteins. *Retrovirology* 1, 12. doi:10.1186/1742-4690-1-12

Aldrich, J. V., and McLaughlin, J. P. (2012). Opioid peptides: Potential for drug development. *Drug Discov. Today Technol.* 9, e23–e31. doi:10.1016/j.ddtec.2011.07.007

Al-Tebrineh, J., Mihali, T. K., Pomati, F., and Neilan, B. A. (2010). Detection of saxitoxin-producing cyanobacteria and Anabaena circinalis in environmental water blooms by quantitative PCR. *Appl. Environ. Microbiol.* 76, 7836–7842. doi:10.1128/aem.00174-10

Andersson, D. I., and Hughes, D. (2010). Antibiotic resistance and its cost: Is it possible to reverse resistance? *Nat. Rev. Microbiol.* 8, 260–271. doi:10.1038/nrmicro2319

Andreevskaia, S. N., Chernousova, L. N., Smirnova, T. G., Larionova, E. E., and Kuz'min, A. V. (2006). *Mycobacterium tuberculosis* strain transmission caused by migratory processes in the Russian Federation (in case of populational migration from the Caucasian Region to Moscow and the Moscow Region). *Probl. Tuberk. Bolezn. Legk.* 1. 29–35.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Ashida, H., Mimuro, H., Ogawa, M., Kobayashi, T., Sanada, T., Kim, M., et al. (2011). Cell death and infection: A double-edged sword for host and pathogen survival. *J. Cell Biol.* 195, 931–942. doi:10.1083/jcb.201108081

Awad, M. M., Johanesen, P. A., Carter, G. P., Rose, E., and Lyras, D. (2014). *Clostridium difficile* virulence factors: Insights into an anaerobic spore-forming pathogen. *Gut Microbes* 5, 579–593. doi:10.4161/19490976.2014.969632

Ayyavoo, V., Mahboubi, A., Mahalingam, S., Ramalingam, R., Kudchodkar, S., Williams, W. V., et al. (1997). HIV-1 Vpr suppresses immune activation and apoptosis through regulation of nuclear factor κB. *Nat. Med.* 3, 1117–1123. doi:10.1038/nm1097-1117

Bakour, S., Sankar, S. A., Rathored, J., Biagini, P., Raoult, D., and Fournier, P. E. (2016). Identification of virulence factors and antibiotic resistance markers using bacterial genomics. *Future Microbiol.* 11, 455–466. doi:10.2217/fmb.15.149

Barth, H., Aktories, K., Popoff, M. R., and Stiles, B. G. (2004). Binary bacterial toxins: Biochemistry, biology, and applications of common Clostridium and Bacillus proteins. *Microbiol. Mol. Biol. Rev.* 68, 373–402. doi:10.1128/mmbr.68.3.373-402.2004

Bartra, S. S., Styer, K. L., O'Bryant, D. M., Nilles, M. L., Hinnebusch, B. J., Aballay, A., et al. (2008). Resistance of *Yersinia pestis* to complement-dependent killing is mediated by the Ail outer membrane protein. *Infect. Immun.* 76, 612–622. doi:10.1128/iai.01125-07

Basso, P., Ragno, M., Elsen, S., Reboud, E., Golovkine, G., Bouillot, S., et al. (2017). *Pseudomonas aeruginosa* pore-forming exolysin and type IV pili cooperate to induce host cell lysis. *MBio* 8, e02250–16. doi:10.1128/mbio.02250-16

Benfield, A. P., Goodey, N. M., Phillips, L. T., and Martin, S. F. (2007). Structural studies examining the substrate specificity profiles of PC-PLC(Bc) protein variants. *Arch. Biochem. Biophys.* 460, 41–47. doi:10.1016/j.abb.2007.01.023

Bernard, S. C., Simpson, N., Join-Lambert, O., Federici, C., Laran-Chich, M. P., Maissa, N., et al. (2014). Pathogenic Neisseria meningitidis utilizes CD147 for vascular colonization. *Nat. Med.* 20, 725–731. doi:10.1038/nm.3563

Blair, J. M., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. (2015). Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* 13, 42–51. doi:10.1038/nrmicro3380

Blight, K. J., Kolykhalov, A. A., and Rice, C. M. (2000). Efficient initiation of HCV RNA replication in cell culture. *Science* 290, 1972–1974. doi:10.1126/science.290.5498.1972

Borzenkov, V. M., Pomerantsev, A. P., and Ashmarin, I. P. (1993). The additive synthesis of a regulatory peptide *in vivo*: The administration of a vaccinal francisella tularensis strain that produces beta-endorphin. *Biull. Eksp. Biol. Med.* 116, 151–153.

Borzenkov, V. M., Pomerantsev, A. P., Pomerantseva, O. M., and Ashmarin, I. P. (1994). Study of nonpathogenic strains of francisella, brucella and yersinia as producers of recombinant beta-endorphin. *Biull. Eksp. Biol. Med.* 117, 612–615.

Brbic, M., Piskorec, M., Vidulin, V., Krisko, A., Smuc, T., and Supek, F. (2016). The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Res.* 44, 10074–10090. doi:10.1093/nar/gkw964

Burns, D. (2003). *Bacterial protein toxins*. Washington, D.C. ASM Press.

Campos, A., and Vasconcelos, V. (2010). Molecular mechanisms of microcystin toxicity in animal cells. *Int. J. Mol. Sci.* 11, 268–287. doi:10.3390/ijms11010268

Casadevall, A., and Pirofski, L. A. (2003). The damage-response framework of microbial pathogenesis. *Nat. Rev. Microbiol.* 1, 17–24. doi:10.1038/nrmicro732

Casilag, F., Lorenz, A., Krueger, J., Klawonn, F., Weiss, S., and Haussler, S. (2016). The LasB elastase of *Pseudomonas aeruginosa* acts in concert with alkaline protease AprA to prevent flagellin-mediated immune recognition. *Infect. Immun.* 84, 162–171. doi:10.1128/iai.00939-15

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46, D633–D639. doi:10.1093/nar/gkx935

Cello, J., Paul, A. V., and Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. *Science* 297, 1016–1018. doi:10.1126/science.1072266

Chambers, H. F. (1997). Methicillin resistance in staphylococci: Molecular and biochemical basis and clinical implications. *Clin. Microbiol. Rev.* 10, 781–791. doi:10.1128/cmr.10.4.781

Chapman, D. A., Tcherepanov, V., Upton, C., and Dixon, L. K. (2008). Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. *J. Gen. Virol.* 89, 397–408. doi:10.1099/vir.0.83343-0

Chen, Z., Franco, C. F., Baptista, R. P., Cabral, J. M., Coelho, A. V., Rodrigues, C. J., Jr., et al. (2007). Purification and identification of cutinases from Colletotrichum kahawae and Colletotrichum gloeosporioides. *Appl. Microbiol. Biotechnol.* 73, 1306–1313. doi:10.1007/s00253-006-0605-1

Chen, N., Bellone, C. J., Schriewer, J., Owens, G., Fredrickson, T., Parker, S., et al. (2011). Poxvirus interleukin-4 expression overcomes inherent resistance and vaccine-induced immunity: Pathogenesis, prophylaxis, and antiviral therapy. *Virology* 409, 328–337. doi:10.1016/j.virol.2010.10.021

Colf, L. A. (2016). Preparing for nontraditional biothreats. *Health Secur.* 14, 7–12. doi:10.1089/hs.2015.0045

Cook, J. D., and Lee, J. E. (2013). The secret life of viral entry glycoproteins: Moonlighting in immune evasion. *PLoS Pathog.* 9, e1003258. doi:10.1371/journal.ppat.1003258

Cornelis, G. R. (2000). Molecular and cell biology aspects of plague. *Proc. Natl. Acad. Sci. U. S. A.* 97, 8778–8783. doi:10.1073/pnas.97.16.8778

Cosentino, S., Voldby Larsen, M., Moller Aarestrup, F., and Lund, O. (2013). PathogenFinder--distinguishing friend from foe using bacterial whole genome sequence data. *PLoS One* 8, e77302. doi:10.1371/journal.pone.0077302

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., et al. (2005). The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature* 434, 980–986. doi:10.1038/nature03449

Dickers, K. J., Bradberry, S. M., Rice, P., Griffiths, G. D., and Vale, J. A. (2003). Abrin poisoning. *Toxicol. Rev.* 22, 137–142. doi:10.2165/00139709-200322030-00002

Dobrindt, U. (2005). (Patho-)Genomics of *Escherichia coli*. *Int. J. Med. Microbiol.* 295, 357–371. doi:10.1016/j.ijmm.2005.07.009

Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* 356, 133–145. doi:10.1098/rstb.2000.0758

Dudak, F. C., Boyaci, I. H., and Orner, B. P. (2011). The discovery of small-molecule mimicking peptides through phage display. *Molecules* 16, 774–789. doi:10.3390/molecules16010774

Dulon, S., Leduc, D., Cottrell, G. S., D'Alayer, J., Hansen, K. K., Bunnett, N. W., et al. (2005). *Pseudomonas aeruginosa* elastase disables proteinase-activated receptor 2 in respiratory epithelial cells. *Am. J. Respir. Cell Mol. Biol.* 32, 411–419. doi:10.1165/rcmb.2004-0274oc

Ebrahimi, M., Aghagolzadeh, P., Shamabadi, N., Tahmasebi, A., Alsharifi, M., Adelson, D. L., et al. (2014). Understanding the underlying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein. *PLoS One* 9, e96984. doi:10.1371/journal.pone.0096984

EMBL-EBI (2019). Toxin activity. Available at: https://www.ebi.ac.uk/QuickGO/term/GO:0090729 (Accessed November 18, 2019).

Erana, H., and Castilla, J. (2016). The architecture of prions: How understanding would provide new therapeutic insights. *Swiss Med. Wkly.* 146, w14354. doi:10.4414/smw.2016.14354

Espinosa Angarica, V., Angulo, A., Giner, A., Losilla, G., Ventura, S., and Sancho, J. (2014). PrionScan: An online database of predicted prion domains in complete proteomes. *BMC Genomics* 15, 102. doi:10.1186/1471-2164-15-102

Esvelt, K. M., Smidler, A. L., Catteruccia, F., and Church, G. M. (2014). Concerning RNA-guided gene drives for the alteration of wild populations. *Elife* 3, e03401. doi:10.7554/elife.03401

Federal Registar (2022). Screening framework guidance for providers and users of synthetic oligonucleotides. Available at: https://www.federalregister.gov/documents/2022/04/29/2022-09210/screening-framework-guidance-for-providers-and-users-of-synthetic-oligonucleotides.

Finlay, B. B. (2005). Bacterial virulence strategies that utilize Rho GTPases. *Curr. Top. Microbiol. Immunol.* 291, 1–10. doi:10.1007/3-540-27511-8_1

Flores-Diaz, M., and Alape-Giron, A. (2003). Role of *Clostridium perfringens* phospholipase C in the pathogenesis of gas gangrene. *Toxicon* 42, 979–986. doi:10.1016/j.toxicon.2003.11.013

Forrellad, M. A., Klepp, L. I., Gioffre, A., Sabio y Garcia, J., Morbidoni, H. R., de la Paz Santangelo, M., et al. (2013). Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* 4, 3–66. doi:10.4161/viru.22329

Fournier, P. E., Richet, H., and Weinstein, R. A. (2006). The epidemiology and control of Acinetobacter baumannii in health care facilities. *Clin. Infect. Dis.* 42, 692–699. doi:10.1086/500202

Francica, J. R., Varela-Rohena, A., Medvec, A., Plesa, G., Riley, J. L., and Bates, P. (2010). Steric shielding of surface epitopes and impaired immune recognition induced by the ebola virus glycoprotein. *PLoS Pathog.* 6, e1001098. doi:10.1371/journal.ppat.1001098

Francis, J. W., Brown, R. H., Jr., Figueiredo, D., Remington, M. P., Castillo, O., Schwarzschild, M. A., et al. (2000). Enhancement of diphtheria toxin potency by replacement of the receptor binding domain with tetanus toxin C-fragment: A potential vector for delivering heterologous proteins to neurons. *J. Neurochem.* 74, 2528–2536. doi:10.1046/j.1471-4159.2000.0742528.x

Friebe, S., van der Goot, F. G., and Burgi, J. (2016). The ins and outs of anthrax toxin. *Toxins (Basel)* 8, 69. doi:10.3390/toxins8030069

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., et al. (2010). Smpdb: The small molecule pathway database. *Nucleic Acids Res.* 38, D480–D487. doi:10.1093/nar/gkp1002

Galanie, S., Thodey, K., Trenchard, I. J., Filsinger Interrante, M., and Smolke, C. D. (2015). Complete biosynthesis of opioids in yeast. *Science* 349, 1095–1100. doi:10.1126/science.aac9373

Galyov, E. E., Brett, P. J., and DeShazer, D. (2010). Molecular insights into Burkholderia pseudomallei and Burkholderia mallei pathogenesis. *Annu. Rev. Microbiol.* 64, 495–517. doi:10.1146/annurev.micro.112408.134030

Gautam, A., Chaudhary, K., Singh, S., Joshi, A., Anand, P., Tuknait, A., et al. (2014). Hemolytik: A database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.* 42, D444–D449. doi:10.1093/nar/gkt1008

Geisinger, E., and Isberg, R. R. (2017). Interplay between antibiotic resistance and virulence during disease promoted by multidrug-resistant bacteria. *J. Infect. Dis.* 215, S9–S17. doi:10.1093/infdis/jiw402

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi:10.1126/science.1190719

Gilmour, M. W., Graham, M., Reimer, A., and Van Domselaar, G. (2013). Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics* 16, 25–30. doi:10.1159/000342709

Godbold Gd, K. A., LeSassier, D. S., Treangen, T. J., and Ternus, K. L. (2021). Categorizing sequences of concern by function to better assess mechanisms of microbial pathogenesis. *Infect. Immun.* 90, e0033421. doi:10.1128/iai.00334-21

Gold, J. A., Hoshino, Y., Jones, M. B., Hoshino, S., Nolan, A., and Weiden, M. D. (2007). Exogenous interferon-alpha and interferon-gamma increase lethality of murine inhalational anthrax. *PLoS One* 2, e736. doi:10.1371/journal.pone.0000736

Goldberg, J. B. (2010). Why is *Pseudomonas aeruginosa* a pathogen? F1000. *F1000 Biol. Rep.* 2, 29. doi:10.3410/b2-29

Goldman, A. S. (2000). Back to basics: Host responses to infection. *Pediatr. Rev.* 21, 342–349. doi:10.1542/pir.21.10.342

Green, E. R., and Mecsas, J. (2016). Bacterial secretion systems: An overview. *Microbiol. Spectr.* 4, 1–32. doi:10.1128/microbiolspec.vmbf-0012-2015

Haas, B., and Grenier, D. (2018). Understanding the virulence of Streptococcus suis: A veterinary, medical, and economic challenge. *Med. Maladies Infect.* 48, 159–166. doi:10.1016/j.medmal.2017.10.001

Hamada, S., Kawabata, S., and Nakagawa, I. (2015). Molecular and genomic characterization of pathogenic traits of group A Streptococcus pyogenes. *Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.* 91, 539–559. doi:10.2183/pjab.91.539

Harbi, D., Parthiban, M., Gendoo, D. M., Ehsani, S., Kumar, M., Schmitt-Ulms, G., et al. (2012). PrionHome: A database of prions and other sequences relevant to prion phenomena. *PLoS One* 7, e31785. doi:10.1371/journal.pone.0031785

Haschek, W., and Voss, K. (2013). *Rousseaux's handbook of toxicologic pathology*. Amsterdam, Netherlands: Elsevier.

Hauswedell, H., Singer, J., and Reinert, K. (2014). Lambda: The local aligner for massive biological data. *Bioinformatics* 30, i349–55. doi:10.1093/bioinformatics/btu439

Headquarters Department of the Army (2018) Nuclear and chemical weapons and materiel chemical surety. Available at: https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN3125_AR50-6_WEB_FINAL.pdf (Accessed April 16, 2018).

Herfst, S., Schrauwen, E. J., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., et al. (2012). Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* 336, 1534–1541. doi:10.1126/science.1213362

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: An open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D455. doi:10.1093/nar/gkh052

Herpfer, I., Katzev, M., Feige, B., Fiebich, B. L., Voderholzer, U., and Lieb, K. (2007). Effects of substance P on memory and mood in healthy male subjects. *Hum. Psychopharmacol. Clin. Exp.* 22, 567–573. doi:10.1002/hup.876

Ho, N. K., Henry, A. C., Johnson-Henry, K., and Sherman, P. M. (2013). Pathogenicity, host responses and implications for management of enterohemorrhagic *Escherichia coli* O157:H7 infection. *Can. J. Gastroenterology* 27, 281–285. doi:10.1155/2013/138673

Holden, M. T., Titball, R. W., Peacock, S. J., Cerdeno-Tarraga, A. M., Atkins, T., Crossman, L. C., et al. (2004). Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. *Proc. Natl. Acad. Sci. U. S. A.* 101, 14240–14245. doi:10.1073/pnas.0403302101

Hu, P., Yang, M., Zhang, A., Wu, J., Chen, B., Hua, Y., et al. (2011). Complete genome sequence of Streptococcus suis serotype 3 strain ST3. *J. Bacteriol.* 193, 3428–3429. doi:10.1128/jb.05018-11

Huang, W. J., Chen, W. W., and Zhang, X. (2015). Prions mediated neurodegenerative disorders. *Eur. Rev. Med. Pharmacol. Sci.* 19, 4028–4034.

Hudson, C. M., Lau, B. Y., and Williams, K. P. (2015). Islander: A database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res.* 43, D48–D53. doi:10.1093/nar/gku1072

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., et al. (2011). ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* 39, D576–D582. doi:10.1093/nar/gkq901

Hutchison, C. A., 3rd, Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351, aad6253. doi:10.1126/science.aad6253

Hwang, I. Y., Koh, E., Wong, A., March, J. C., Bentley, W. E., Lee, Y. S., et al. (2017). Engineered probiotic *Escherichia coli* can eliminate and prevent *Pseudomonas aeruginosa* gut infection in animal models. *Nat. Commun.* 8, 15028. doi:10.1038/ncomms15028

IARPA. Broad Agency Announcement (2016). Functional genomic and computational assessment of threats (fun GCAT). IARPA-BAA-16-08. Available at: https://viterbi.usc.edu/links/webuploads/Functional%20Genomic%20and%20Computational%20Assessment%20of%20Threats%20(Fun%20GCAT)%20IARPA-BAA-16-08.pdf.

Ilina, E. N., Shitikov, E. A., Ikryannikova, L. N., Alekseev, D. G., Kamashev, D. E., Malakhova, M. V., et al. (2013). Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia. *PLoS One* 8, e56577. doi:10.1371/journal.pone.0056577

Inoshima, I., Inoshima, N., Wilke, G. A., Powers, M. E., Frank, K. M., Wang, Y., et al. (2011). A *Staphylococcus aureus* pore-forming toxin subverts the activity of ADAM10 to cause lethal infection in mice. *Nat. Med.* 17, 1310–1314. doi:10.1038/nm.2451

International Gene Synthesis Consortium (2017) Harmonized screening protocol v2.0 gene sequence & customer screening to promote biosecurity. Available at: https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf (Accessed November 19, 2017).

Ireton, K. (2013). Molecular mechanisms of cell-cell spread of intracellular bacterial pathogens. *Open Biol.* 3, 130079. doi:10.1098/rsob.130079

Izard, T., Tran Van Nhieu, G., and Bois, P. R. (2006). Shigella applies molecular mimicry to subvert vinculin and invade host cells. *J. Cell Biol.* 175, 465–475. doi:10.1083/jcb.200605091

Jackson, R. J., Ramsay, A. J., Christensen, C. D., Beaton, S., Hall, D. F., and Ramshaw, I. A. (2001). Expression of mouse interleukin-4 by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mousepox. *J. Virol.* 75, 1205–1210. doi:10.1128/jvi.75.3.1205-1210.2001

Jacoby, G. A., and Munoz-Price, L. S. (2005). The new beta-lactamases. *N. Engl. J. Med. Overseas. Ed.* 352, 380–391. doi:10.1056/nejmra041359

Jal, S., and Khora, S. S. (2015). An overview on the origin and production of tetrodotoxin, a potent neurotoxin. *J. Appl. Microbiol.* 119, 907–916. doi:10.1111/jam.12896

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). Card 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi:10.1093/nar/gkw1004

Jirschitzka, J., Schmidt, G. W., Reichelt, M., Schneider, B., Gershenzon, J., and D'Auria, J. C. (2012). Plant tropane alkaloid biosynthesis evolved independently in the Solanaceae and Erythroxylaceae. *Proc. Natl. Acad. Sci. U. S. A.* 109, 10304–10309. doi:10.1073/pnas.1200473109

Jorgensen, R., Purdy, A. E., Fieldhouse, R. J., Kimber, M. S., Bartlett, D. H., and Merrill, A. R. (2008). Cholix toxin, a novel ADP-ribosylating factor from *Vibrio cholerae*. *J. Biol. Chem.* 283, 10671–10678. doi:10.1074/jbc.m710008200

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., et al. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432. doi:10.1093/nar/gki072

Jungo, F., Bougueleret, L., Xenarios, I., and Poux, S. (2012). The UniProtKB/Swiss-prot tox-prot program: A central hub of integrated venom protein data. *Toxicon* 60, 551–557. doi:10.1016/j.toxicon.2012.03.010

Kastin, A. (2013). *Handbook of biologically active peptides*. Amsterdam, Netherlands: Elsevier.

Kaur, P., Chakraborti, A., and Asea, A. (2010). Enteroaggregative *Escherichia coli*: An emerging enteric food borne pathogen. *Interdiscip. Perspect. Infect. Dis.* 2010, 1–10. doi:10.1155/2010/254159

Kempf, M., and Rolain, J. M. (2012). Emergence of resistance to carbapenems in acinetobacter baumannii in europe: Clinical impact and therapeutic options. *Int. J. Antimicrob. Agents* 39, 105–114. doi:10.1016/j.ijantimicag.2011.10.004

Kerr, P. J., Perkins, H. D., Inglis, B., Stagg, R., McLaughlin, E., Collins, S. V., et al. (2004). Expression of rabbit IL-4 by recombinant myxoma viruses enhances virulence and overcomes genetic resistance to myxomatosis. *Virology* 324, 117–128. doi:10.1016/j.virol.2004.02.031

Khairil Anuar, I. N. A., Banerjee, A., Keeble, A. H., Carella, A., Nikov, G. I., and Howarth, M. (2019). Spy&Go purification of SpyTag-proteins using pseudo-SpyCatcher to access an oligomerization toolbox. *Nat. Commun.* 10, 1734. doi:10.1038/s41467-019-09678-w

Kiessling, K. (1986). Biochemical mechanism of action of mycotoxins. *Pure Appl. Chem.* 58, 327–338. doi:10.1351/pac198658020327

Klein, A. M., Vaissiere, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., et al. (2007). Importance of pollinators in changing landscapes for world crops. *Proc. R. Soc. B* 274, 303–313. doi:10.1098/rspb.2006.3721

Knodler, L. A., Celli, J., and Finlay, B. B. (2001). Pathogenic trickery: Deception of host cell processes. *Nat. Rev. Mol. Cell Biol.* 2, 578–588. doi:10.1038/35085062

Kohler, C. D., and Dobrindt, U. (2011). What defines extraintestinal pathogenic *Escherichia coli*? *Int. J. Med. Microbiol.* 301, 642–647. doi:10.1016/j.ijmm.2011.09.006

Korbsrisate, S., Tomaras, A. P., Damnin, S., Ckumdee, J., Srinon, V., Lengwehasatit, I., et al. (2007). Characterization of two distinct phospholipase C enzymes from Burkholderia pseudomallei. *Microbiology* 153, 1907–1915. doi:10.1099/mic.0.2006/003004-0

Kurupati, P., Turner, C. E., Tziona, I., Lawrenson, R. A., Alam, F. M., Nohadani, M., et al. (2010). Chemokine-cleaving Streptococcus pyogenes protease SpyCEP is necessary and sufficient for bacterial dissemination within soft tissues and the respiratory tract. *Mol. Microbiol.* 76, 1387–1397. doi:10.1111/j.1365-2958.2010.07065.x

Kuzmenkov, A. I., Krylov, N. A., Chugunov, A. O., Grishin, E. V., and Vassilevski, A. A. (2016). Kalium: A database of potassium channel toxins from scorpion venom. *Database (Oxford)* 2016, baw056. doi:10.1093/database/baw056

Lago, J., Rodriguez, L. P., Blanco, L., Vieites, J. M., and Cabado, A. G. (2015). Tetrodotoxin, an extremely potent marine neurotoxin: Distribution, toxicity, origin and therapeutical uses. *Mar. Drugs* 13, 6384–6406. doi:10.3390/md13106384

Lamkanfi, M., and Dixit, V. M. (2010). Manipulation of host cell death pathways during microbial infections. *Cell Host Microbe* 8, 44–54. doi:10.1016/j.chom.2010.06.007

Legname, G., Baskakov, I. V., Nguyen, H. O., Riesner, D., Cohen, F. E., DeArmond, S. J., et al. (2004). Synthetic mammalian prions. *Science* 305, 673–676. doi:10.1126/science.1100195

Lewis, R., Dutertre, S., Vetter, I., and Christie, M. (2012). Conus venom peptide pharmacology. *Pharmacol. Rev.* 64, 259–298. doi:10.1124/pr.111.005322

Li, Q., Zhang, C., Chen, H., Xue, J., Guo, X., Liang, M., et al. (2018). BioPepDB: An integrated data platform for food-derived bioactive peptides. *Int. J. Food Sci. Nutr.* 69, 963–968. doi:10.1080/09637486.2018.1446916

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). Vfdb 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 47, D687–D692. doi:10.1093/nar/gky1080

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., et al. (2019). ICEberg 2.0: An updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* 47, D660–D665. doi:10.1093/nar/gky1123

Lu, T., Yao, B., and Zhang, C. (2012). Dfvf: Database of fungal virulence factors. *Database (Oxford)*, 2012 bas032. doi:10.1093/database/bas032

Lu, Q. F., Cao, D. M., Su, L. L., Li, S. B., Ye, G. B., Zhu, X. Y., et al. (2019). Genus-wide comparative genomics analysis of Neisseria to identify new genes associated with pathogenicity and niche adaptation of Neisseria pathogens. *Int. J. Genomics* 2019, 1–19. doi:10.1155/2019/6015730

Luo, N., Pereira, S., Sahin, O., Lin, J., Huang, S., Michel, L., et al. (2005). Enhanced *in vivo* fitness of fluoroquinolone-resistant Campylobacter jejuni in the absence of antibiotic selection pressure. *Proc. Natl. Acad. Sci. U. S. A.* 102, 541–546. doi:10.1073/pnas.0408966102

Luo, G., Ibrahim, A. S., Spellberg, B., Nobile, C. J., Mitchell, A. P., and Fu, Y. (2010). Candida albicans Hyr1p confers resistance to neutrophil killing and is a potential vaccine target. *J. Infect. Dis.* 201, 1718–1728. doi:10.1086/652407

Magarlamov, T. Y., Melnikova, D. I., and Chernyshev, A. V. (2017). Tetrodotoxin-producing bacteria: Detection, distribution and migration of the toxin in aquatic systems. *Toxins (Basel)* 9, 166. doi:10.3390/toxins9050166

Mahenthiralingam, E., and Vandamme, P. (2005). Taxonomy and pathogenesis of the Burkholderia cepacia complex. *Chron. Respir. Dis.* 2, 209–217. doi:10.1191/1479972305cd053ra

Mathur, D., Prakash, S., Anand, P., Kaur, H., Agrawal, P., Mehta, A., et al. (2016). PEPlife: A repository of the half-life of peptides. *Sci. Rep.* 6, 36617. doi:10.1038/srep36617

Memisevic, V., Kumar, K., Cheng, L., Zavaljevski, N., DeShazer, D., Wallqvist, A., et al. (2014). DBSecSys: A database of Burkholderia mallei secretion systems. *BMC Bioinforma.* 15, 244. doi:10.1186/1471-2105-15-244

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016). Linking virus genomes with host taxonomy. *Viruses* 8, 66. doi:10.3390/v8030066

Mitchell, A. M., and Mitchell, T. J. (2010). Streptococcus pneumoniae: Virulence factors and variation. *Clin. Microbiol. Infect.* 16, 411–418. doi:10.1111/j.1469-0691.2010.03183.x

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi:10.1093/nar/gky1100

Mueller, M., Grauschopf, U., Maier, T., Glockshuber, R., and Ban, N. (2009). The structure of a cytolytic alpha-helical toxin pore reveals its assembly mechanism. *Nature* 459, 726–730. doi:10.1038/nature08026

Nakagawa, A., Matsumura, E., Koyanagi, T., Katayama, T., Kawano, N., Yoshimatsu, K., et al. (2016). Total biosynthesis of opiates by stepwise fermentation using engineered *Escherichia coli*. *Nat. Commun.* 7, 10390. doi:10.1038/ncomms10390

National Cancer Institute (2019). Aflatoxins. Available at: https://www.cancer.gov/about-cancer/causes-prevention/risk/substances/aflatoxins (Accessed October 10, 2019).

NCBI (2019). Batch Entrez. Available at: https://www.ncbi.nlm.nih.gov/sites/batchentrez (Accessed November 18, 2019).

Nesic, D., and Stebbins, C. E. (2005). Mechanisms of assembly and cellular interactions for the bacterial genotoxin CDT. *PLoS Pathog.* 1, e28. doi:10.1371/journal.ppat.0010028

Newby, D. E., Sciberras, D. G., Ferro, C. J., Gertz, B. J., Sommerville, D., Majumdar, A., et al. (1999). Substance P-induced vasodilatation is mediated by the neurokinin type 1 receptor but does not contribute to basal vascular tone in man. *Br. J. Clin. Pharmacol.* 48, 336–344. doi:10.1046/j.1365-2125.1999.00017.x

Nguyen, Y., and Sperandio, V. (2012). Enterohemorrhagic *E. coli* (EHEC) pathogenesis. *Front. Cell. Infect. Microbiol.* 2, 90. doi:10.3389/fcimb.2012.00090

Nielsen, S. D., Beverly, R. L., Qu, Y., and Dallas, D. C. (2017). Milk bioactive peptide database: A comprehensive database of milk protein-derived bioactive peptides and novel visualization. *Food Chem. x.* 232, 673–682. doi:10.1016/j.foodchem.2017.04.056

Niu, C., Yu, D., Wang, Y., Ren, H., Jin, Y., Zhou, W., et al. (2013). Common and pathogen-specific virulence factors are different in function and structure. *Virulence* 4, 473–482. doi:10.4161/viru.25730

Noor Uddin, G. M., Larsen, M. H., Christensen, H., Aarestrup, F. M., Phu, T. M., and Dalsgaard, A. (2015). Identification and antimicrobial resistance of bacteria isolated from probiotic products used in shrimp culture. *PLoS One* 10, e0132338. doi:10.1371/journal.pone.0132338

Noyce, R. S., Lederman, S., and Evans, D. H. (2018). Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. *PLoS One* 13, e0188453. doi:10.1371/journal.pone.0188453

Nuermberger, E. L., and Bishai, W. R. (2004). Antibiotic resistance in Streptococcus pneumoniae: What does the future hold? *Clin. Infect. Dis.* 38 (4), S363–S371. doi:10.1086/382696

O'Brien, A. D., Tesh, V. L., Donohue-Rolfe, A., Jackson, M. P., Olsnes, S., Sandvig, K., et al. (1992). Shiga toxin: Biochemistry, genetics, mode of action, and role in pathogenesis. *Curr. Top. Microbiol. Immunol.* 180, 65–94. doi:10.1007/978-3-642-77238-2_4

Oldfield, L. M., Grzesik, P., Voorhies, A. A., Alperovich, N., MacMath, D., Najera, C. D., et al. (2017). Genome-wide engineering of an infectious clone of herpes simplex virus type 1 using synthetic genomics assembly methods. *Proc. Natl. Acad. Sci. U. S. A.* 114, E8885–E8894. doi:10.1073/pnas.1700534114

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541. doi:10.1093/nar/gky1079

Park, H., Valencia-Gallardo, C., Sharff, A., Tran Van Nhieu, G., and Izard, T. (2011). Novel vinculin binding site of the IpaA invasin of Shigella. *J. Biol. Chem.* 286, 23214–23221. doi:10.1074/jbc.m110.184283

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinforma.* 3, 1. doi:10.1002/0471250953.bi0301s42

Pedraza, L. A., Bautista, J., and Uribe-Velez, D. (2018). Seed-born Burkholderia glumae infects rice seedling and maintains bacterial population during vegetative and reproductive growth stage. *Plant Pathol. J.* 34, 393–402. doi:10.5423/ppj.oa.02.2018.0030

Pineda, S. S., Chaumeil, P. A., Kunert, A., Kaas, Q., Thang, M. W. C., Le, L., et al. (2018). ArachnoServer 3.0: An online resource for automated discovery, analysis and annotation of spider toxins. *Bioinformatics* 34, 1074–1076. doi:10.1093/bioinformatics/btx661

Plano, G. V., and Schesser, K. (2013). The *Yersinia pestis* type III secretion system: Expression, assembly and role in the evasion of host defenses. *Immunol. Res.* 57, 237–245. doi:10.1007/s12026-013-8454-3

Poulos, J., and Farnia, A. (2015). *Production of cannabidiolic acid in yeast*. US10093949B2.

Prasanna, A. N., and Mehra, S. (2013). Comparative phylogenomics of pathogenic and non-pathogenic mycobacterium. *PLoS One* 8, e71248. doi:10.1371/journal.pone.0071248

Prusiner, S. B. (1998). Prions. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13363–13383. doi:10.1073/pnas.95.23.13363

Raetz, C. R., and Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* 71, 635–700. doi:10.1146/annurev.biochem.71.110601.135414

Raimbault, B., Cointet, J. P., and Joly, P. B. (2016). Mapping the emergence of synthetic biology. *PLoS One* 11, e0161522. doi:10.1371/journal.pone.0161522

Rasool, S., Hussain, T., Khan, S. M., Zehra, A., Tahreem, S., and Kakroo, A. M. (2017). Toxins of *Clostridium perfringens* as virulence factors in animal diseases. *J. Pharmacogn. Phytochemistry* 6, 2155–2164.

Raymond, B., Young, J. C., Pallett, M., Endres, R. G., Clements, A., and Frankel, G. (2013). Subversion of trafficking, apoptosis, and innate immunity by type III secretion system effectors. *Trends Microbiol.* 21, 430–441. doi:10.1016/j.tim.2013.06.008

Reactome (2019). Reactome. Available at: https://reactome.org/ (Accessed November 18, 2019).

Reboud, E., Basso, P., Maillard, A. P., Huber, P., and Attree, I. (2017). Exolysin shapes the virulence of *Pseudomonas aeruginosa* clonal outliers. *Toxins (Basel)* 9, 364. doi:10.3390/toxins9110364

RefSeq (2019). Growth statistics. Available at: https://www.ncbi.nlm.nih.gov/refseq/statistics/2019.

Reperant, L. A., and Osterhaus, A. (2017). AIDS, avian flu, SARS, MERS, ebola, Zika. What next? *Vaccine* 35, 4470–4474. doi:10.1016/j.vaccine.2017.04.082

Roly, Z. Y., Hakim, M. A., Zahan, A. S., Hossain, M. M., and Reza, M. A. (2015). ISOB: A database of indigenous snake species of Bangladesh with respective known venom composition. *Bioinformation* 11, 107–114. doi:10.6026/97320630011107

Romani, B., and Engelbrecht, S. (2009). Human immunodeficiency virus type 1 vpr: Functions and molecular interactions. *J. Gen. Virol.* 90, 1795–1805. doi:10.1099/vir.0.011726-0

Rooijakkers, S. H., van Kessel, K. P., and van Strijp, J. A. (2005). Staphylococcal innate immune evasion. *Trends Microbiol.* 13, 596–601. doi:10.1016/j.tim.2005.10.002

Roux, D., Danilchanka, O., Guillard, T., Cattoir, V., Aschard, H., Fu, Y., et al. (2015). Fitness cost of antibiotic susceptibility during bacterial infection. *Sci. Transl. Med.* 7, 297ra114. doi:10.1126/scitranslmed.aab1621

Rudel, T., Scheurerpflug, I., and Meyer, T. F. (1995). Neisseria PilC protein identified as type-4 pilus tip-located adhesin. *Nature* 373, 357–359. doi:10.1038/373357a0

Sarovich, D. S., Price, E. P., Webb, J. R., Ward, L. M., Voutsinos, M. Y., Tuanyok, A., et al. (2014). Variable virulence factors in Burkholderia pseudomallei (melioidosis) associated with human disease. *PLoS One* 9, e91682. doi:10.1371/journal.pone.0091682

Sassetti, C. M., and Rubin, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12989–12994. doi:10.1073/pnas.2134250100

Sayers, S., Li, L., Ong, E., Deng, S., Fu, G., Lin, Y., et al. (2019). Victors: A web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res.* 47, D693–D700. doi:10.1093/nar/gky999

Schmid-Hempel, P. (2009). Immune defence, parasite evasion strategies and their relevance for 'macroscopic phenomena' such as virulence. *Phil. Trans. R. Soc. B* 364, 85–98. doi:10.1098/rstb.2008.0157

Schmitt, C. K., Meysick, K. C., and O'Brien, A. D. (1999). Bacterial toxins: Friends or foes? *Emerg. Infect. Dis.* 5, 224–234. doi:10.3201/eid0502.990206

Schuelke, T. A., Wu, G., Westbrook, A., Woeste, K., Plachetzki, D. C., Broders, K., et al. (2017). Comparative genomics of pathogenic and nonpathogenic beetle-vectored fungi in the genus Geosmithia. *Genome Biol. Evol.* 9, 3312–3327. doi:10.1093/gbe/evx242

Segura, M., Fittipaldi, N., Calzas, C., and Gottschalk, M. (2017). Critical Streptococcus suis virulence factors: Are they all really critical? *Trends Microbiol.* 25, 585–599. doi:10.1016/j.tim.2017.02.005

Serpinskii, O. I., Kochneva, G. V., Urmanov, I., Sivolobova, G. F., and Riabchikova, E. I. (1996). Construction of recombinant variants or orthopoxviruses by inserting foreign genes into intragenic region of viral genome. *Mol. Biol.* 30, 1055–1065.

Shah, P. S., Link, N., Jang, G. M., Sharp, P. P., Zhu, T., Swaney, D. L., et al. (2018). Comparative flavivirus-host protein interaction mapping reveals mechanisms of dengue and Zika virus pathogenesis. *Cell* 175, 1931–1945. doi:10.1016/j.cell.2018.11.028

Shames, S. R., and Finlay, B. B. (2010). Breaking the stereotype: Virulence factor-mediated protection of host cells in bacterial pathogenesis. *PLoS Pathog.* 6, e1001057. doi:10.1371/journal.ppat.1001057

Shames, S. R., Deng, W., Guttman, J. A., de Hoog, C. L., Li, Y., Hardwidge, P. R., et al. (2010). The pathogenic *E. coli* type III effector EspZ interacts with host CD98 and facilitates host cell prosurvival signalling. *Cell. Microbiol.* 12, 1322–1339. doi:10.1111/j.1462-5822.2010.01470.x

Shaver, C. M., and Hauser, A. R. (2004). Relative contributions of *Pseudomonas aeruginosa* ExoU, ExoS, and ExoT to virulence in the lung. *Infect. Immun.* 72, 6969–6977. doi:10.1128/iai.72.12.6969-6977.2004

Shen, Y., Chen, L., Wang, M., Lin, D., Liang, Z., Song, P., et al. (2017). Flagellar hooks and hook protein FlgE participate in host microbe interactions at immunological level. *Sci. Rep.* 7, 1433. doi:10.1038/s41598-017-01619-1

Skurnik, D., Roux, D., Cattoir, V., Danilchanka, O., Lu, X., Yoder-Himes, D. R., et al. (2013). Enhanced *in vivo* fitness of carbapenem-resistant oprD mutants of *Pseudomonas aeruginosa* revealed through high-throughput sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 110, 20747–20752. doi:10.1073/pnas.1221552110

Slusarczyk, A. L., Lin, A., and Weiss, R. (2012). Foundations for the design and implementation of synthetic genetic circuits. *Nat. Rev. Genet.* 13, 406–420. doi:10.1038/nrg3227

Smatti, M. K., Cyprian, F. S., Nasrallah, G. K., Al Thani, A. A., Almishal, R. O., and Yassine, H. M. (2019). Viruses and autoimmunity: A review on the potential interaction and molecular mechanisms. *Viruses* 11, 762. doi:10.3390/v11080762

Smedley, J. G., 3rd, Fisher, D. J., Sayeed, S., Chakrabarti, G., and McClane, B. A. (2004). The enteric toxins of *Clostridium perfringens*. *Rev. Physiol. Biochem. Pharmacol.* 152, 183–204. doi:10.1007/s10254-004-0036-2

Smith, H. O., Hutchison, C. A., 3rd, Pfannkoch, C., and Venter, J. C. (2003). Generating a synthetic genome by whole genome assembly: φX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15440–15445. doi:10.1073/pnas.2237126100

Stapleton, P. D., and Taylor, P. W. (2002). Methicillin resistance in *Staphylococcus aureus*: Mechanisms and modulation. *Sci. Prog.* 85, 57–72. doi:10.3184/003685002783238870

Stebbins, C. E., and Galan, J. E. (2001). Structural mimicry in bacterial virulence. *Nature* 412, 701–705. doi:10.1038/35089000

Straus, M. R., and Whittaker, G. R. (2017). A peptide-based approach to evaluate the adaptability of influenza A virus to humans based on its hemagglutinin proteolytic cleavage site. *PLoS One* 12, e0174827. doi:10.1371/journal.pone.0174827

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt, C. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. doi:10.1093/bioinformatics/btu739

Sweet, C. R., Conlon, J., Golenbock, D. T., Goguen, J., and Silverman, N. (2007). YopJ targets TRAF proteins to inhibit TLR-mediated NF-kappaB, MAPK and IRF3 signal transduction. *Cell. Microbiol.* 9, 2700–2715. doi:10.1111/j.1462-5822.2007.00990.x

Sweigard, J. A., Chumley, F. G., and Valent, B. (1992). Cloning and analysis of CUT1, a cutinase gene from Magnaporthe grisea. *Molec. Gen. Genet.* 232, 174–182. doi:10.1007/bf00279994

Tehel, A., Vu, Q., Bigot, D., Gogol-Doring, A., Koch, P., Jenkins, C., et al. (2019). The two prevalent genotypes of an emerging infectious disease, deformed wing virus, cause equally low pupal mortality and equally high wing deformities in host honey bees. *Viruses* 11, 114. doi:10.3390/v11020114

Telling, G. C., Scott, M., Mastrianni, J., Gabizon, R., Torchia, M., Cohen, F. E., et al. (1995). Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell* 83, 79–90. doi:10.1016/0092-8674(95)90236-8

The, C. (2019). Gene ontology, the gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338.

Tillett, D., Dittmann, E., Erhard, M., von Dohren, H., Borner, T., and Neilan, B. A. (2000). Structural organization of microcystin biosynthesis in microcystis aeruginosa PCC7806: An integrated peptide-polyketide synthetase system. *Chem. Biol.* 7, 753–764. doi:10.1016/s1074-5521(00)00021-1

Tsang, T. M., Felek, S., and Krukonis, E. S. (2010). Ail binding to fibronectin facilitates *Yersinia pestis* binding to host cells and Yop delivery. *Infect. Immun.* 78, 3358–3368. doi:10.1128/iai.00238-10

U.S. Department of Health and Human Services (2020). *Biosafety in microbiological and biomedical laboratories*. Sixth Edition. Washington DC. https://www.cdc.gov/labs/pdf/CDC-BiosafetyMicrobiologicalBiomedicalLaboratories-2020-P.pdf.

U.S. Department of Justice. Drug Enforcement Agency. Diversion Control Division (2019) Title 21 United States code (USC) controlled substances act. Available at: https://www.deadiversion.usdoj.gov/21cfr/21usc/index.html (Accessed Nover 18, 2019).

UniProt. (2019) UniRef. Available at: https://www.uniprot.org/help/uniref (Accessed November 18, 2019).

UniProt, C. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

UniProt (2019). Animal toxin annotation project. Available at: https://www.uniprot.org/program/Toxins.

United States Department of Agriculture Economic Research Service (2022) Farming and farm income. Available at: https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/farming-and-farm-income/(Accessed 8 2022).

United States Drug Enforcement Administration (2019). Drug scheduling. Available at: https://www.dea.gov/drug-scheduling (Accessed October 10, 2019).

Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., et al. (2017). PHI-Base: A new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.* 45, D604–D610. doi:10.1093/nar/gkw1089

US Department of Health and Human Services (2017). Framework for guiding funding decisions about proposed research involving enhanced potential pandemic pathogens. Available at: https://www.phe.gov/s3/dualuse/Documents/p3co.pdf.

US Department of Health and Human Services (2022). Screening framework guidance for providers of synthetic double-stranded DNA. Available at: https://www.phe.gov/Preparedness/legal/guidance/syndna/Pages/default.aspx.

Usmani, S. S., Bedi, G., Samuel, J. S., Singh, S., Kalra, S., Kumar, P., et al. (2017). THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS One* 12, e0181748. doi:10.1371/journal.pone.0181748

Uzzau, S., and Fasano, A. (2000). Cross-talk between enteric pathogens and the intestine. *Cell. Microbiol.* 2, 83–89. doi:10.1046/j.1462-5822.2000.00041.x

van Der Most, R. G., Murali-Krishna, K., Ahmed, R., and Strauss, J. H. (2000). Chimeric yellow fever/dengue virus as a candidate dengue vaccine: Quantitation of the dengue virus-specific CD8 T-cell response. *J. Virol.* 74, 8094–8101. doi:10.1128/jvi.74.17.8094-8101.2000

Velmurugan, K., Chen, B., Miller, J. L., Azogue, S., Gurses, S., Hsu, T., et al. (2007). *Mycobacterium tuberculosis* nuoG is a virulence gene that inhibits apoptosis of infected host cells. *PLoS Pathog.* 3, e110. doi:10.1371/journal.ppat.0030110

Vickers, C., and Small, I. (2018) The synthetic biology revolution is now – here's what that means. Available at: https://phys.org/news/2018-09-synthetic-biology-revolution.html2018.

Vieira, A., Silva, D. N., Varzea, V., Paulo, O. S., and Batista, D. (2019). Genome-wide signatures of selection in Colletotrichum kahawae reveal candidate genes potentially involved in pathogenicity and aggressiveness. *Front. Microbiol.* 10, 1374. doi:10.3389/fmicb.2019.01374

Vila, J., Marti, S., and Sanchez-Cespedes, J. (2007). Porins, efflux pumps and multidrug resistance in Acinetobacter baumannii. *J. Antimicrob. Chemother.* 59, 1210–1215. doi:10.1093/jac/dkl509

ViralZone (2019). Viralzone news. Available at: https://viralzone.expasy.org/ (Accessed November 18, 2019).

Visiello, R., Colombo, S., and Carretto, E. (2016). *Chapter 3 - Bacillus cereus hemolysins and other virulence factors, the diverse faces of Bacillus cereus*. Amsterdam, Netherlands: Elsevier, 35–44.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2005). String: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–D437. doi:10.1093/nar/gki005

Wang, J., Yin, T., Xiao, X., He, D., Xue, Z., Jiang, X., et al. (2018). StraPep: A structure database of bioactive peptides. *Database (Oxford)* 2018, bay038. doi:10.1093/database/bay038

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al. (2017). Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* 45, D535–D542. doi:10.1093/nar/gkw1017

Welch, R. A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 17020–17024. doi:10.1073/pnas.252529799

Wells, G. A., Scott, A. C., Johnson, C. T., Gunning, R. F., Hancock, R. D., Jeffrey, M., et al. (1987). A novel progressive spongiform encephalopathy in cattle. *Vet. Rec.* 121, 419–420. doi:10.1136/vr.121.18.419

Whitworth, T., Popov, V. L., Yu, X. J., Walker, D. H., and Bouyer, D. H. (2005). Expression of the *Rickettsia prowazekii* pld or tlyC gene in *Salmonella enterica* serovar Typhimurium mediates phagosomal escape. *Infect. Immun.* 73, 6668–6673. doi:10.1128/iai.73.10.6668-6673.2005

Wilesmith, J. W. (1994). Bovine spongiform encephalopathy and related diseases: An epidemiological overview. *N. Z. Vet. J.* 42, 1–8. doi:10.1080/00480169.1994.35774

Will, R. G., Ironside, J. W., Zeidler, M., Cousens, S. N., Estibeiro, K., Alperovitch, A., et al. (1996). A new variant of Creutzfeldt-Jakob disease in the UK. *Lancet* 347, 921–925. doi:10.1016/s0140-6736(96)91412-9

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. (2016). Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Res.* 44, D646–D653. doi:10.1093/nar/gkv1227

Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., et al. (2015). T3DB: The toxic exposome database. *Nucleic Acids Res.* 43, D928–D934. doi:10.1093/nar/gku1004

Wong, G., Kobinger, G. P., and Qiu, X. (2014). Characterization of host immune responses in Ebola virus infections. *Expert Rev. Clin. Immunol.* 10, 781–790. doi:10.1586/1744666x.2014.908705

Xiong, Z., Jiang, Y., Qi, D., Lu, H., Yang, F., Yang, J., et al. (2009). Complete genome sequence of the extremophilic Bacillus cereus strain Q1 with industrial applications. *J. Bacteriol.* 191, 1120–1121. doi:10.1128/jb.01629-08

Xu, S. X., and McCormick, J. K. (2012). Staphylococcal superantigens in colonization and disease. *Front. Cell. Infect. Microbiol.* 2, 52. doi:10.3389/fcimb.2012.00052

Xu, X., Liu, W., Tian, S., Wang, W., Qi, Q., Jiang, P., et al. (2018). Petroleum hydrocarbon-degrading bacteria for the remediation of oil pollution under aerobic conditions: A perspective analysis. *Front. Microbiol.* 9, 2885. doi:10.3389/fmicb.2018.02885

Yongkiettrakul, S., Maneerat, K., Arechanajan, B., Malila, Y., Srimanote, P., Gottschalk, M., et al. (2019). Antimicrobial susceptibility of Streptococcus suis isolated from diseased pigs, asymptomatic pigs, and human patients in Thailand. *BMC Vet. Res.* 15, 5. doi:10.1186/s12917-018-1732-5

Yoon, S. H., Park, Y. K., and Kim, J. F. (2015). PAIDB v2.0: Exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res.* 43, D624–D630. doi:10.1093/nar/gku985

Zakham, F., Aouane, O., Ussery, D., Benjouad, A., and Ennaji, M. M. (2012). Computational genomics-proteomics and Phylogeny analysis of twenty one mycobacterial genomes (Tuberculosis & non Tuberculosis strains). *Microb. Inf. Exp.* 2, 7. doi:10.1186/2042-5783-2-7

Zaluga, J., Stragier, P., Baeyen, S., Haegeman, A., Van Vaerenbergh, J., Maes, M., et al. (2014). Comparative genome analysis of pathogenic and non-pathogenic Clavibacter strains reveals adaptations to their lifestyle. *BMC Genomics* 15, 392. doi:10.1186/1471-2164-15-392

Zamyatnin, A. A., Borchikov, A. S., Vladimirov, M. G., and Voronina, O. L. (2006). The EROP-Moscow oligopeptide database. *Nucleic Acids Res.* 34, D261–D266. doi:10.1093/nar/gkj008

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., et al. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res.* 45, D466–D474. doi:10.1093/nar/gkw857

Zhang, H. (2003). Lethality in mice infected with recombinant vaccinia virus expressing hepatitis C virus core protein. *Hepatobiliary Pancreat. Dis. Int.* 2, 374–382.

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., and Slezak, T. (2007). MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* 35, D391–D394. doi:10.1093/nar/gkl791