Standardizing Functional Annotation of Sequences Involved in Microbial Pathogenesis to Better Understand Proteins and Enable Bioinformatic Applications

Gene D. Godbold • Jody B. G. Proescher • Matthew B. Scholz • Anthony D. Kappell • Todd J. Treangen • Krista L. Ternus

Introduction

Available annotations of sequences enabling microbial pathogenesis are inconsistent, lack mechanistic specificity, do not specify host targets, and are frequently contaminated with host sequences. This makes their use challenging for computational applications. To improve understanding of the functions of sequences enabling microbial pathogenesis and the ability to recognize such sequences computationally, we reviewed thousands of papers in microbial pathogenesis, annotating more than 3000 virulence factors from more than 105 bacterial species, 85 viruses, and 25 eukaryotic pathogens. This review informed the development of a compact (~30 term) controlled vocabulary: Functions of Sequences of Concern (FunSoCs) to describe their role and consequences in pathogenesis. We distinguish proteins that act directly on host molecules and those that act indirectly, by affecting only other parasite molecules. FunSoC machine learning assignments are available through our open-source SeqScreen software: https://gitlab.com/treangenlab/seqscreen.

In addition, we assisted with the development of the Pathogenesis Gene Ontology (PathGO), led by researchers at the Johns Hopkins University Applied Physics Lab. PathGO is a set of ~170 terms specific to microbial pathogenesis: https://github.com/jhuapl-bio/pathogenesis-gene-ontology. Among other advantages, PathGO terms allow a more specific marking of host target pathways that FunSoCs must neglect to remain compact. Our dataset is annotated with both FunSoCs and PathGO terms to improve our understanding of virulence factors and the threats they pose in a bioengineering, gain-of-function scenario.

Functions of Sequences of Concern (FunSoCs)

SEQUENCES OF CONCERN THAT OPERATE **DIRECTLY**

Damaging SoCs

- Cytotoxic
- Degrade Tissue
- Disable Organ
- Induce Inflammation

Other Direct-Acting SoCs

- Adherence to host cell
- Dissemination in host
- Host cell invasion
- Movement in host cell
- Niche-creation in host cell

Host Cellular Processes Manipulated by SoCs

- Apoptosis
- Autophagy/Xenophagy
- Cell cycle
- Cytoskeleton dynamics
- Endomembrane dynamics
- Small GTPase dynamics
- Transcription
- Translation
- Ubiquitination

Immune-Subverting SoCs

- Suppress host immune signaling by disrupting:
  - JAK-STAT
  - MAPK
  - NFκB
  - RIG-I
  - Protein kinase R
  - STING
  - TNF receptor-associated factors
  - Toll-like receptor binding/ activation
- Resist host phagocytosis
- Resist host complement-killing (serum resistance)
- Resist host antimicrobial peptide
- Resist host oxidative killing
- Counter host immunoglobulin
- Defeat host chemokine or cytokine
- Inhibit host antigen presentation
- Immunomodulation

SEQUENCES OF CONCERN THAT OPERATE **INDIRECTLY**

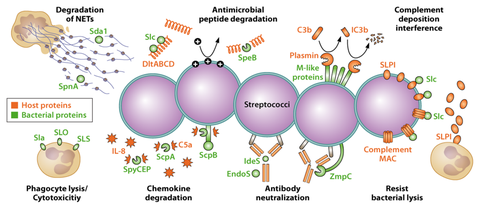
- Bacterial capsule biosynthetic enzymes
- Passive immune evasion enzymes (modify parasite structures)
- Secretion system components and chaperones that transport parasite effectors
- Toxin biosynthetic enzymes (e.g. mycotoxin synthases)
- Transcription factors influencing SoCs

Ranking Of Sequence Biothreats

1. **Direct > Indirect**
2. **Multi-function > Single Function**
3. **Damaging > Disseminating > Adhering** or **Invading > Within-Cell Movement** or **Niche-Creation**
4. **Innate Immune-Subverting ≥ Damaging?**

Hypothesis

Innate Immune-subverting SoCs make a host susceptible to the encoding parasite.



Examples of innate immune subversion by Streptococcal effectors. Host phagocytes are debilitated by streptolysin O (SLO), streptolysin S (SLS), and secreted phospholipase A2 (Sla), Neutrophil extracellular traps (NETs) are countered by the Sda1 and SpnA nucleases. Antimicrobial peptides are inactivated by the secreted streptococcal inhibitor of complement (Sic) and SpeB proteases. M-like proteins bind host factor H and plasminogen/ plasmin which inactivate host complement components to protect the bacterium. Sic protects streptococci from phagocytosis by neutrophils, resists host complement membrane attack complex (MAC), and counters the antibacterial actions of the host secretory leukocyte proteinase inhibitor (SLPI). Host antibodies are destroyed by membrane-associated ZmpC and the secreted IdeS proteases and inactivated by sugar-cleaving EndoS. The group B streptococcus C5a peptidase ScpB is a serine protease and surface invasin that reduces the neutrophil response and bacterial clearance by cutting the chemoattractant C5a. The streptococcal complement protector ScpA helps the bacterium resist phagocytosis and also inactivates C5a. SpyCEP eliminates the neutrophil chemoattractant IL-8 and other chemokines. The figure depicts SoCs found in both GAS and GBS for illustrative purposes, but they would *NOT* naturally be found together.

How many microbial pathogens for humans are there?[1-3]

~600 **Fungi** ~600 **Bacteria**

50 **RNA Viruses** 250 **DNA Viruses**

~60 **Protozoa**

Sample Annotations Of Sequences Of Concern

We have annotated Sequences of Concern (SoCs) from 105 bacterial, 58 viral, seven protozoal, and four fungal species pathogenic for humans.
More if pathogens of livestock and crop plants are included.

Sample Annotations from Bacterial, Viral, Fungal, and Protozoal Pathogens

SoC, Organism	FunSoCs	PathGO terms
LasB, Pseudomonas aeruginosa	Resist host complement; Resist host antimicrobial peptide; Resist host oxidative killing; Counter host cytokine; Resist other host immune effector; Induce inflammation; Degrade tissue; Disable organ;	PATHGO:0000271 (mediates resistance to oxidative killing in another organism); PATHGO:0000353 (modulates reactive oxygen species levels in another organism); PATHGO:0000100 (mediates resistance to complement system in another organism); PATHGO:0000104 (disrupts antimicrobial peptide binding in another organism); PATHGO:0000363 (suppresses pro-inflammatory cytokine activity in another organism); PATHGO:0000214 (modifies tight junction or adherens junction in another organism); PATHGO:0000358 (mediates release of cell from extracellular matrix in another organism);
IbpA, Histophilus somni	Manipulate host small GTPase; Manipulate host cytoskeleton dynamics; Adherence to another organism; Resist host phagocytosis; Resist host complement; Counter host immunoglobulin; Cytotoxicity;	PATHGO:0000355 (mediates deactivation of small GTPase in another organism); PATHGO:0000216 (mediates filamentous actin depolymerization in another organism); PATHGO:0000211 (mediates binding to the cell surface in another organism); PATHGO:0000232 (suppresses phagocytosis in another organism); PATHGO:0000257 (mediates immunoglobulin neutralization in another organism); PATHGO:0000100 (mediates resistance to complement system in another organism);
IpaB, Shigella flexneri	Manipulate host cell cycle; Secretion system component; Adherence to another organism; Host invasion; Suppress host immune signaling; Induce inflammation; Cytotoxicity;	PATHGO:0000152 (induces cell cycle arrest in cell of another organism); PATHGO:0000110 (mediates secretion of protein effector); PATHGO:0000234 (mediates binding to integrin in another organism); PATHGO:0000368 (mediates host cell invasion by microbe); PATHGO:0000220 (suppresses inflammatory cytokine release in another organism); PATHGO:0000284 (mediates binding to cholesterol in another organism); PATHGO:0000033 (mediates pore formation in another organism);
TcdA, Clostridioides difficile	Manipulate host small GTPase; Manipulate host cytoskeleton dynamics; Adherence to another organism; Host invasion; Induce inflammation; Degrade tissue;	PATHGO:0000285 (mediates carbohydrate-derivative binding in another organism); PATHGO:0000273 (mediates glycosaminoglycan- or proteoglycan-binding in another organism); PATHGO:0000072 (mediates binding to cell surface glycoprotein in another organism); PATHGO:0000214 (modifies tight junction or adherens junction in another organism); PATHGO:0000369 (mediates cell invasion by macromolecule from another organism); PATHGO:0000355 (mediates deactivation of small GTPase in another organism); PATHGO:0000214 (modifies tight junction or adherens junction in another organism); PATHGO:0000216 (mediates filamentous actin depolymerization in another organism); PATHGO:0000162 (disrupts epithelium in another organism);
NS1, influenza virus	Manipulate host transcription; Manipulate host translation; Manipulate host ubiquitin dynamics; Manipulate host regulated cell death; Suppress host immune signaling; Resist other host immune effector; Suppress antigen presentation;	PATHGO:0000326 (modulates transcription in another organism); PATHGO:0000006 (modulates protein synthesis in another organism); PATHGO:0000325 (modulates ubiquitin dynamics in another organism); PATHGO:0000352 (disrupts TRIM/TRIM-like signaling in another organism); PATHGO:0000334 (suppresses apoptosis in another organism); PATHGO:0000220 (suppresses inflammatory cytokine release in another organism); PATHGO:0000080 (suppresses dendritic cell activation in another organism); PATHGO:0000312 (mediates concealment of foreign nucleic acid in another organism);
E1A, human adenovirus	Manipulate host transcription; Manipulate host cell cycle; Manipulate host ubiquitin dynamics; Manipulate host regulated cell death; Suppress host immune signaling; Suppress antigen presentation	PATHGO:0000325 (modulates ubiquitin dynamics in another organism); PATHGO:0000326 (modulates transcription in another organism); PATHGO:0000335 (induces apoptosis in another organism); PATHGO:0000300 (disrupts STING signaling in another organism); PATHGO:0000308 (disrupts antigen presentation in another organism);
NSs, Rift Valley fever virus	Manipulate host transcription; Manipulate host cell cycle; Manipulate host ubiquitin dynamics; Manipulate host cytoskeleton dynamics; Suppress host immune signaling; Resist other host immune effector;	PATHGO:0000326 (modulates transcription in another organism); PATHGO:0000152 (induces cell cycle arrest in cell of another organism); PATHGO:0000325 (modulates ubiquitin dynamics in another organism); PATHGO:0000028 (modulates cytoskeleton in another organism); PATHGO:0000214 (modifies tight junction or adherens junction in another organism); PATHGO:0000382 (suppresses interferon signaling in another organism); PATHGO:0000304 (disrupts PKR activity in another organism);
Alp1, Neosartorya fumigata	Resists host complement; Counter host immunoglobulin; Degrade tissue; Disable organ;	PATHGO:0000100 (mediates resistance to complement system in another organism); PATHGO:0000257 (mediates immunoglobulin neutralization in another organism); PATHGO:0000226 (disrupts extracellular matrix in another organism);
ROP18/VIR3, Toxoplasma gondii	Manipulate host ubiquitin dynamics; Manipulate host programmed cell death; Suppress host immune signaling	PATHGO:0000325 (modulates ubiquitin dynamics in another organism); PATHGO:0000295 (suppresses NFκB signaling in another organism); PATHGO:0000334 (suppresses apoptosis in another organism); PATHGO:0000352 (disrupts TRIM/TRIM-like signaling in another organism);

Conclusions

A Dataset of Annotated SoCs Can Bring Clarity to Dual-Use Research of Concern (DURC) for Researchers, Regulators, and Funding Agencies.

Function of Sequence of Concern (FunSoC)	CASE 1, SoC transferred to other pathogen or CASE 2, SoC altered for enhancement of original pathogen	CASE 3, SoC transferred to nonpathogen
Damaging	Could enhance the harmful consequences of the agent;	Might enable the nonpathogen to have harmful consequences
Immune-Subverting	Enhances the harmful consequences of the agent; Disrupts immunity or the effectiveness of an immunization against the agent; Alters the host range or tropism of the agent; Enhances the susceptibility of a host population to the agent;	Might enable the nonpathogen to have harmful consequences; Might enable the nonpathogen to infect novel hosts; Might enhance the susceptibility of a host population to the agent;
Attachment Protein/ Adhesin	Alters the host range or tropism of the agent; Enhances the susceptibility of a host population to the agent;	Probably none
Fusion Protein/ Invasin	Alters the host range or tropism of the agent; Enhances the susceptibility of a host population to the agent;	Probably none
Dissemination	Enhances the harmful consequences of the agent; Increases the transmissibility or the ability to disseminate the agent; Enhances the susceptibility of a host population to the agent;	Probably none

References

1. Woolhouse M, Gaunt E. (2007) Ecological origins of novel human pathogens. Crit Rev Microbiol 33:231–242. 10.1080/10408410701647560.
2. Brown GD, Denning DW, Levitz SM. (2012) Tackling human fungal infections. Science 336:647. 10.1126/science.1222236.
3. Woolhouse MEJ, Brierley L. (2018) Epidemiological characteristics of human-infective RNA viruses. Sci Data 5:180017. 10.1038/sdata.2018.17.
4. Godbold, GD, Kappell, AD, LeSassier, DS, Treangen, TJ, and Ternus, KL (2022) Categorizing Sequences of Concern by Function To Better Assess Mechanisms of Microbial Pathogenesis. Infect Immun 90, e0033421. doi: 10.1128/IAI.00334-21.
5. Balaji, A, Kille, B, Kappell, AD, Godbold GD, Diep, M, Elworth RAL, Qian, Z, Albin, D, Nasko, DJ, Shah, N, Pop, M, Segarra S, Ternus, KL, Treangen TJ (2022) SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. Genome Biol 23: 133. doi: 10.1186/s13059-022-02695-x.

Acknowledgements

Gene D. Godbold, Jody B. G. Proescher, Matthew B. Scholz, Anthony D. Kappell, Todd J. Treangen, and Krista L. Ternus were partially supported by the Fun GCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under federal award no. W911NF-17-2-0089.

Signature Science, LLC
8329 North Mopac Expressway
Austin, TX 78759

Signature Science, LLC
1670 Discovery Drive
Charlottesville, VA 22911

Johns Hopkins University Applied Physics
Assymmetric Operations Sector
Laurel, MD 20723

Rice University
Department of Computer Science
Houston, TX 77005

signature science® LLC