Annotation of Microbial Proteins to Identify Pathogenic Functions

Matthew Scholz

PROPRIETARY INFORMATION

Acknowledgements

Krista Ternus
Principle Investigator

Gene D. Godbold

Anthony D. Kappell

Danielle S. LeSassier

Todd Treangen
Principle Investigator

Advait Balaji

Work comprising this talk has been done in partial fulfillment of the goals of:

- **IARPA Fun GCAT Program**
- **IARPA Raven Project**
- **CDC Harvest Variants Project**

All of the co-authors were either fully or 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. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

What Are We Worried About?

- Goals:
  - Look for indications of potential pathogens from sequence data
  - Classify a sequenced genome to determine whether it could be capable of pathogenicity
- Questions:
  - What makes a sequence ‘Dangerous’?
  - What does a given sequence DO?
  - How do we know any of this?

PROPRIETARY INFORMATION

Biothreat Agents – A Good Start

- Select Agent and Toxins (HHS/USDA)
  - 37 Viruses
  - 18 Bacteria
  - Four Fungi/Oomycetes
  - Nine Toxins
  - Not inclusive of all potential biothreats

PROPRIETARY INFORMATION

What's Already Out There?

- Current projects are targeted or incomplete
  - **VFDB**: Not manually curated, missing details.
  - **PHI-Base**: Limited to outlining gene alterations resulting in pathogenicity changes
  - **Victors**: Included based on a KO experiment for the gene
  - **CARD Virulence Ontology**: an undergraduate project
  - **PATRIC**: 10 bacterial features, four of which are 'pathogenic'
  - **UniProt:** Noisy, permissive, inaccurate.
- How do we make a better database?

PROPRIETARY INFORMATION

There are Approximately **1 TRILLION** Microbial Species on Earth

VIRUSES • BACTERIA • FUNGI • PROTOZOA

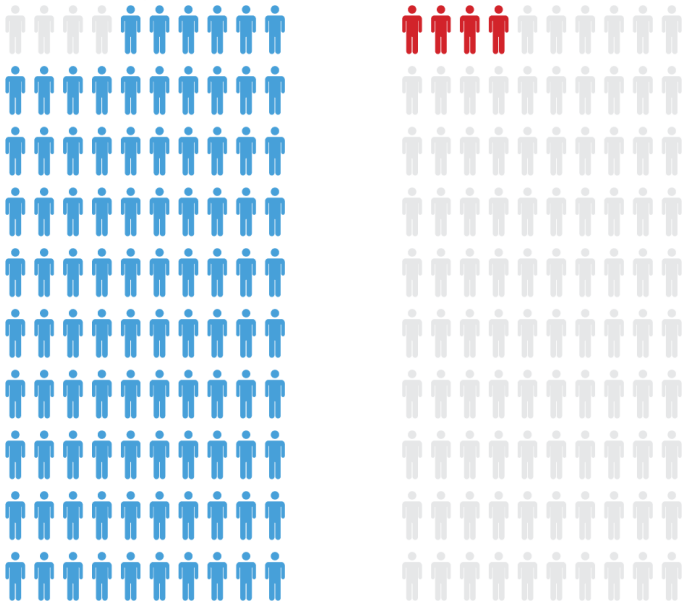
1,300
Viruses, Bacteria, Fungi, Protozoa, and Worms

700
Bacteria and Fungi

MILLIONS
of Microbial Species Can Cause Disease in a Human with Essentially No Immune System

Cause Disease in Humans with Normal Immunity (~96%)

Cause Disease in Immune Compromised Humans (~4%)



SCID
Severe Combined Immunodeficiency (~0.000001%)

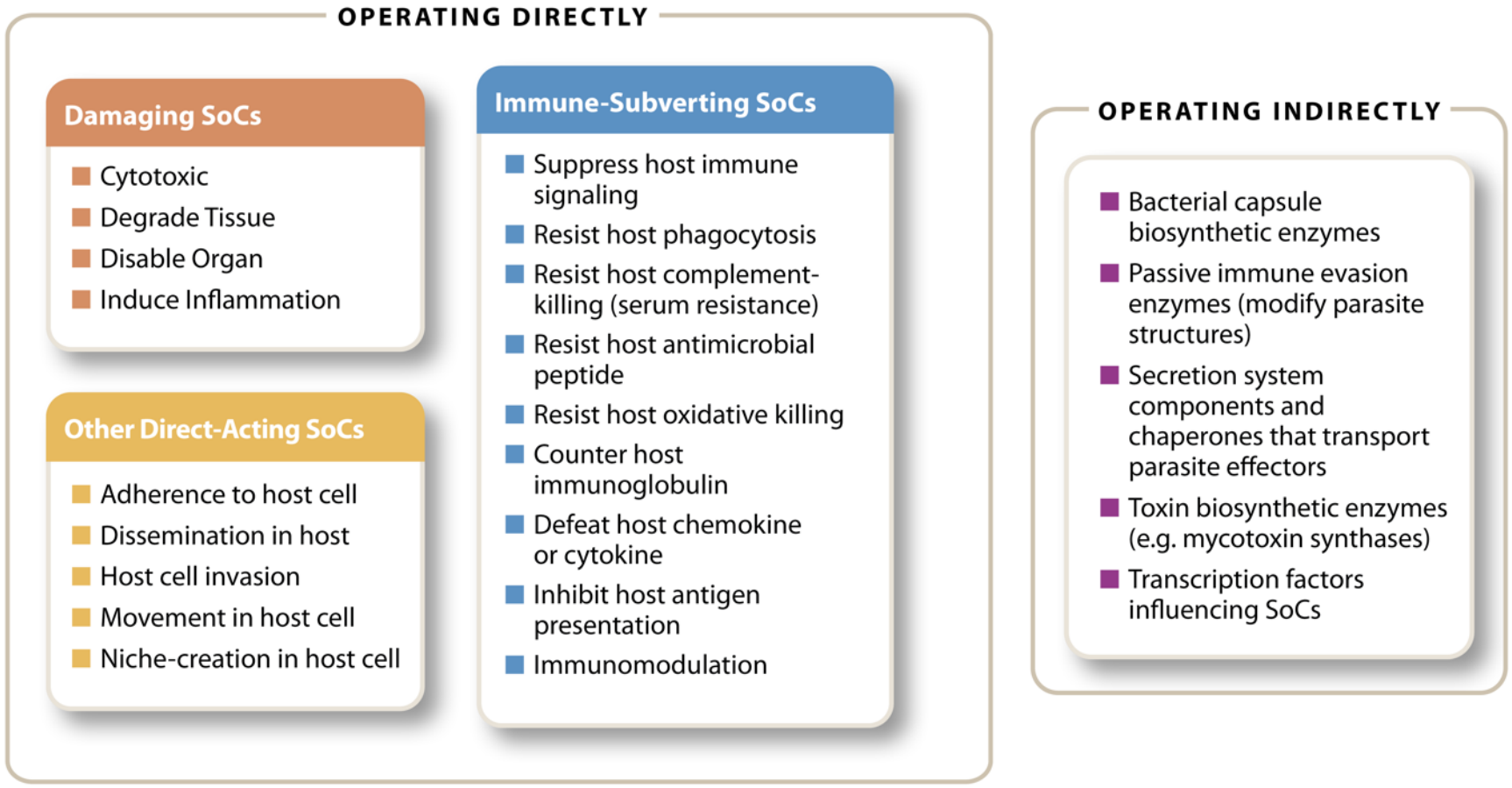
signature science LLC

Sequences of Concern

- SoCs
  - What causes the concern (How do we determine something is linked to pathogenicity)?
    - Presence in pathogens?
    - Annotation in a Database?
  - What is our range of hosts?
    - Humans?
    - Animals?
    - Plants?
- The reason a sequence causes concern is due to its FUNCTION

PROPRIETARY INFORMATION

Functions of Sequences of Concern (FunSoCs)



Godbold *et al.*, 2022: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9119117/

PROPRIETARY INFORMATION

Diving Deeper

- FunSoCs describe what is happening (actions/functions)
  - What is the actual mode of action of an expressed sequence?
  - Is there an identified target for that sequence?
  - Multiple activities lumped together.
- Pathogenesis Gene Ontology (PathGO)
  - Supported development by Johns Hopkins University Applied Physics Laboratory
  - ~170 Unique terms to describe sequences' mode of action in regards to potential to cause harm
  - Hierarchical structure
  - https://github.com/jhuapl-bio/pathogenesis-gene-ontology.

PROPRIETARY INFORMATION

Empty Frameworks are Useless

- Identifying sequences to associate with PathGO terms or FunSoCs requires *manual curation*
  - Analyzed :
    - 100+ Bacterial Species
    - 85 Viruses
    - 25 Eukaryotic pathogens
  - Identified
    - Sequences with published mechanisms associated with one or more terms
    - Identical sequences in other organisms (Amino Acid)

PROPRIETARY INFORMATION

What Are The Applications?

- Stored in MongoDB database
  - >3000 Sequences with associated annotations for FunSoCs and PathGO terms
  - Forms for adding new annotations/altering current annotations
  - Queries
- Incorporation into SeqScreen sequence analysis tool
  - https://gitlab.com/treangenlab/seqscreen
- Eventual incorporation of PathGO terms into Gene Ontology Consortium?
  - Talks underway

PROPRIETARY INFORMATION

Ranking Pathogenicity

- Hierarchy of 'Concern' for sequences (Most to least impact):
  1) Sequences that enable evasion or suppression of immune response.
  2) Sequences with multiple annotated, pathogenic modes of action.
  3) Sequences encoding for functionality allow for dissemination.
  4) Sequences with direct modes of action (damage to cellular membranes).
  5) Sequences which provide indirect action (binding to cells or matrix).
  6) Sequences which allow for intercellular movement or niche formation are of the lowest concern for the purposes of biothreat classification.

PROPRIETARY INFORMATION

What Did We Find?

- Total sequences: 7583
- Total FunSoC terms: 32
- Total PathGO terms: ~140

Most Frequently Annotated FunSoCs	
Adherence to Another Organism	1547
Secretion System Component	1449
Suppress Host Immune Signaling	1162
Host Invasion	1135
Cytotoxicity,Permeabilize Host Cell	1063
Total Annotations	**31184**

Most Frequent PathGO Terms		
PATHGO_0000384	Effector Proteins	538
PATHGO_0000322	Antiobiotic resistance	479
PATHGO_0000110	Protein secretion	262
PATHGO_0000211	Cell to Cell Binding	233
PATHGO_0000337	Toxin synthesis	189
Total Annotations	**10832**	

PROPRIETARY INFORMATION

What CAN We Do with this Information?

- Applied knowledge:
  - Seqscreen annotates hits to describe potential pathogens
  - S2Fast (Gov't Use Rights) can classify the threat LEVEL of a sample
- Extrapolation
  - Can these annotations be extended?
    - UniRef 100 annotated
    - UniRef 90?
    - UniRef 50?

PROPRIETARY INFORMATION

PROPRIETARY INFORMATION

Questions?

PROPRIETARY INFORMATION