Mining Microbial Sequences of Concern:
Using Similarity to Understand Pathogen Risk and Evolution

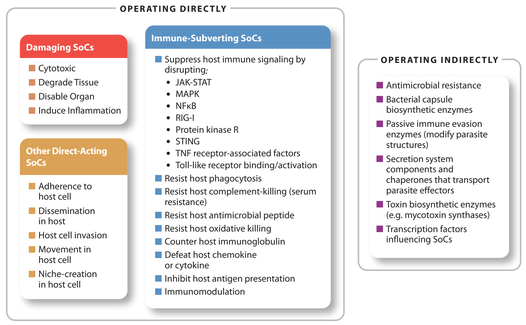
David L. Tyus[1,2] (SIGNATURE SCIENCE and UVA BIOMEDICAL SCIENCES) • **Matthew Scholz**[3] and **Gene D. Godbold**[1] (SIGNATURE SCIENCE)

BACKGROUND

What are Sequences of Concern (SOCs)?

Salmonella, E. coli and Coronavirus are very different organisms but they share notoriety for their disease-causing capabilities. Microbes employ a variety of strategies to survive within hosts but a "pathogen" antagonizes/damages host cells while avoiding immune defense surveillance. This means that microbes in and of themselves aren't the culprits until they gain the capability to create havoc in the body. Genetic determinants that allow for pathogenic activity, called "Sequences of Concern"(SoC) are the difference between innocuous and harmful microbes.

Functions of Sequences of Concern (FunSoCs)



Increasing Surveillance and Understanding of Pathogenic Activity via Sequences of Concern

To keep track of factors that lead pathogenicity it is necessary to categorically separate genetic factors that pathogens "have" from ones that make them pathogens. Our group has worked to create a system of databasing and functional annotation for Sequences of Concern (SoC), sequences that enable microbial pathogenesis. The current database holds 2,750 highly annotated SoCs to facilitate risk monitoring of pathogenic activity. Still, we recognize the potential for many more related, less-annotated SoCs to exist as pathogens have horizontally and vertically passed these virulence factors throughout evolutionary history.

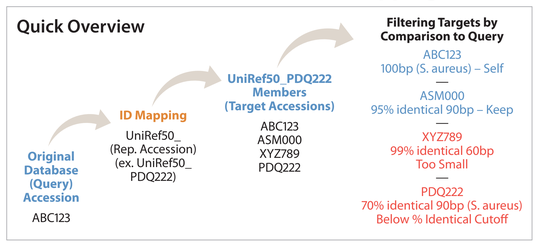
"We seek to expand the number of SoCs we capture by leveraging data on sequence similarity to high-confidence SoCs in our current database."

KEY TERMS

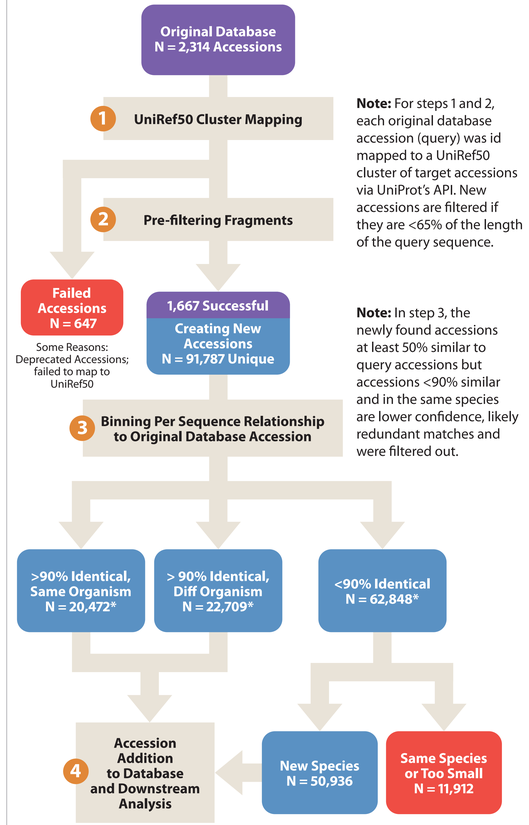
- **UniProt** – Public database of DNA sequences
- **Accession** – Unique identifier for a given sequence (a protein can have multiple accessions associated with it)
- **UniRef50** – UniProt's clustering algorithm that has put accessions into collections based on sequence similarity. Each "UniRef50" cluster has sequences that are at least 50% similar over the 80% of the length of a representative sequence of the collection

METHODS

Expansion of SoC Database Through Automated UniRef50 ID Mapping



The Process of Finding New Potential SoCs



Note: For steps 1 and 2, each original database accession (query) was id mapped to a UniRef50 cluster of target accessions via UniProt's API. New accessions are filtered if they are <65% of the length of the query sequence.

Note: In step 3, the newly found accessions at least 50% similar to query accessions but accessions <90% similar and in the same species are lower confidence, likely redundant matches and were filtered out.

Note: We aim to compare and contextualize sequences of high confidence with the new collection of SoCs with analyses such as MSA and network analysis not only to better estimate the functional capacity of different sequences as virulence factors but to also capture relationships between sequences that point to evolutionary transfer and/or genetic exchange of SoCs.

*The binned accessions are not the sum of the unique accessions as some new accessions are compared to more than one original database accessions as they come from different multiple UniRef50 clusters.

RESULTS OF DATABASE EXPANSION

Represented Proteins

How Would Database Expansion Change the Types of SoCs Represented?

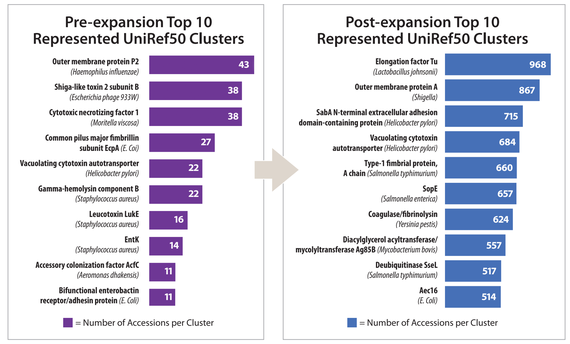


Figure 1: While the original database is varied in the UniRef50 clusters it represents, there is higher representation of accessions of bacteria *E. coli, S. aureus, Streptococci,* and *M. tuberculosis.*

Figure 2: The most represented UniRef50 clusters in the original database are no longer in the expanded database. Even the most represented organisms have changed with *Salmonella* species dominating.

Are the UniRef50 clusters most represented dependent on ...

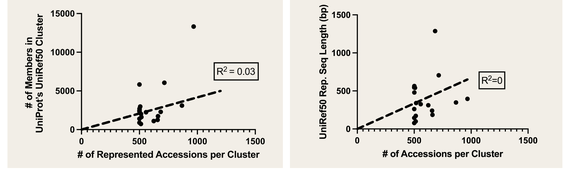


Figure 3: No, the most represented accessions are not the number of accession in UniProt's UniRef50 clusters or the sequence length of the clusters' representatives.

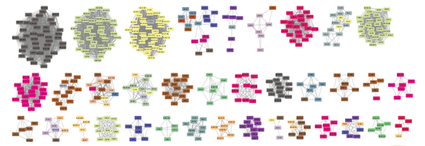
"UniRef50-based expansion of the SoC database fundamentally changes the types of proteins most represented. This change occurs independently of UniRef50 cluster size and sequence length."

Network Analysis

How Can Database Expansion Facilitate Answering Questions Regarding Bacterial Pathogen Evolution and Gene Transfer?

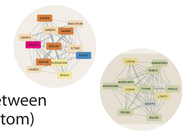
Network analysis with EFI's Enzyme Similarity Tool and visualized with Cytoscape

Pre-expansion (Original Accessions)

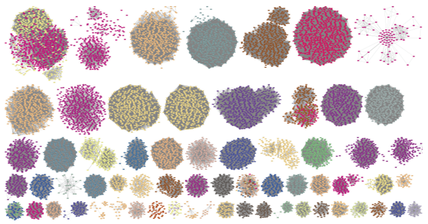


Each node represents an accession. The color designates its UniRef50. Connections are made between sequences >50% identical.

- Most of the accessions in the original database are well organized into uniform UniRef50 cluster networks (top showing a subset of the networks created)
- There are cases where networks are made between accessions of different UniRef50 clusters (bottom)



Post-expansion (Potential New SoCs)



New Network Features

- With the addition of new accessions, there is an increased variety of structure of networks showing clusters within clusters, central nodes, and crosstalk between UniRef50 clusters
- Highly structured networks like the examples above are ideal candidates for downstream analyses such as Multiple Sequence Alignment to as central nodes and detached clusters could give insight to evolutionary changes in SoC-like sequences and potentially point to divergences in functional capabilities

REFERENCES

For more about Sequences of Concern:

Godbold et al. **Categorizing Sequences of Concern by Function To Better Assess Mechanisms of Microbial Pathogenesis.** *Infect Immun.* 2022 May 19;90(5):e0033421. doi: 10.1128/IAI.00334-21. Epub 2021 Nov 15. PMID: 34780277; PMCID: PMC9119117

Godbold et al. **Improved understanding of biorisk for research involving microbial modification using annotated sequences of concern.** *Front Bioeng Biotechnol.* 2023 Apr 25;11:1124100. doi: 10.3389/fbioe.2023.1124100. PMID: 37180048; PMCID: PMC10167326.

EFI Tools (For Network Analysis):

Nils Oberg, Rémi Zallot, and John A. Gerlt, **EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools.** *J Mol Biol* 2023. https://doi.org/10.1016/j.jmb.2023.168018

Cytoscape(For Network Visualization):

Ideker, T. (2003). **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research,* 13(11), 2498–2504.

CONCLUSION and FUTURE DISCUSSION

- The potential number of SoCs increased from ~17,000 sequences to ~80,000 sequences with UniRef50-based searches for similar sequences to our original SoC database
- The types of proteins and organisms most represented will expand in interesting ways as a result of database expansion
- The increased number of sequences per protein allows for more complex structures in network analysis which could signify functional differences following sequence modification through evolutionary history
- An overarching goal for this project is to widen the scope of risk management against SoCs. A more complete database will aid in the surveillance of microbial threats in spite of genotypic variation

[1] **Signature Science, LLC**
1670 Discovery Drive
Charlottesville, VA 22911

[2] **Neuroscience Graduate Program**
University of Virginia
Charlottesville, VA 22901

[3] **Signature Science, LLC**
8329 North Mopac Expressway
Austin, TX 78759