Control Number: 2022-A-86-NGS

Topic 1: Bridging Silos: Exploring mechanisms for collecting and sharing microbial genomic data to foster interoperability

Publishing Title: Fostering Genomic Data Interoperability via MIxS Metadata Standards

Author Block: **L. M. Schriml**[1], E. Eloe-Fadrosh[2], R. Walls[3];
[1]University of Maryland School of Medicine, Baltimore, MD, [2]Joint Genome Institute, Department of Energy, Berkeley, CA, [3]Critical Path Institute, Tucson, AZ.

Abstract Body: Big genomic data integration holds the promise of accessible datasets amenable to ML/AI approaches for knowledge discovery. To achieve this goal, the Genomic Standards Consortium (GSC) fostered the development of a community to build contextual metadata standards, established a suite of genomic checklists and environment-specific packages to enable collection and reporting of standardized metadata. These "Minimum Information about any (X) Sequence" (MIxS) standards support data interoperability and have been expanded and developed across diverse research communities over the past ten years to meet the challenges of big genomic data generation. The MIxS standards further standardize reported metadata through the utilization of ontology standards (e.g. Human Disease Ontology, EnvO, PATO, Uberon, FMA, OBI, GAZ, ChEBI). MIxS standards are broadly utilized across genomic data repositories and resources, including the INSDC's (International Nucleotide Sequence Database Collaboration: NCBI, EMBL-EBI, DDBJ) ENA, GenBank, BioSample and Biosamples repositories along with the National Microbiome Data Collaborative, QIITA/QIIME, CyVerse, JGI's GOLD database and ENA's MGnify resource, thus providing the mechanism for standardized and interoperable contextual data reporting and data reuse. The GSC further supports data interoperability through the establishment of data standardization policies with journals, including GigaScience, Nature's Scientific Data, ISME Journal, BioMed Central's Microbiome and Environmental Microbiome Journals, ASM's mSystems, and Marine Genomics. The GSC is enhancing data interoperability across initiatives through the adoption and collaborative development of biodiversity data exchange formats with the Ocean Biomolecular Observing Network (OBON) and the United Nations Ocean Decade Programme. Teaming up with the Biodiversity Information Standard (TDWG) producing novel mappings between the MIxS standard and Darwin Core standard (DwC) that are being utilized to facilitate data interoperability with biodiversity data including the Global Ocean Observing System's (GOOS) Essential Ocean Variables (EOV's). I will review the development and utilization of standards fostering genomic metadata integration. I will examine methods for improving contextual metadata within and between genomic projects and methods for enhancing interoperability between genomic datasets and sample data repositories.

---PAGEBREAK---

Control Number: 2022-A-89-NGS

Topic 1: Bridging Silos: Exploring mechanisms for collecting and sharing microbial genomic data to foster interoperability

Publishing Title: What Are 'Bad' Sequences? Functional Annotation of Sequences of Concern (SoCs) Enabling Microbial Pathogenesis

Author Block: **G. D. Godbold**[1], J. B. Proescher[2], R. J. Jacak[3], A. D. Kappell[4], T. J. Treangen[5], K. L. Ternus[4]; [1]Signature Science, Charlottesville, VA, [2]Asymmetric Operations Sector, The Johns Hopkins University Applied Physics Laboratory, Baltimore, MD, [3]Research and Exploratory Development Department, The Johns Hopkins University Applied Physics Laboratory, Baltimore, MD, [4]Signature Science, Austin, TX, [5]Department of Computer Science, Rice University, Houston, TX.

Abstract Body: Available annotations of sequences enabling microbial pathogenesis present difficulties including inconsistency, lack of specificity for mechanisms and targets, sometimes too much specificity for viral pathogens, annotations for specific viral proteins assigned instead to the polyproteins from which they are cleaved, and contamination with host sequences. This complicates both human understanding and machine use.To improve our understanding of sequences of concern and their recognition by machines, we manually reviewed thousands of publications in microbial pathogenesis. We annotated more than 2700 virulence factors from more than 100 bacterial species, 85 viruses, and 20 eukaryotic pathogens to develop a compact controlled vocabulary: FunSoCs (Functions of Sequences of Concern). FunSoC assignments are available through our open-source SeqScreen software: https://gitlab.com/treangenlab/seqscreen. In a parallel effort, we assisted the development of the Pathogenesis Gene Ontology (PathGO), led by researchers at the Johns Hopkins University Applied Physics Lab. PathGO is currently a set of ~180 terms describing biological processes comprising microbial pathogenesis: https://github.com/jhuapl-bio/pathogenesis-gene-ontology. The more granular PathGO terms allow the automated recognition of host targets that FunSoCs neglect while FunSoCs capture the organismal consequences of host pathogen interactions that are out of scope for the PathGO terms.The most concerning sequences of concern (SoCs) are toxins and toxic effectors as they damage a host. Next are those sequences that manipulate the host immune system. We have documented over 600 immune-subverting sequences including more than 100 bacterial, viral, and protozoal disruptors of host NF-kappaB signaling. We distinguish between sequences that associate with host targets and those that exert their effects on pathogenesis secondarily by interacting with a parasite molecule. We anticipate that these complementary controlled vocabularies and the proteins annotated using them will be useful for microbial genomics, public health, biosecurity, biosurveillance, and the characterization of new and emerging pathogens.

---PAGEBREAK---

Control Number: 2022-A-111-NGS

Topic 1: Bridging Silos: Exploring mechanisms for collecting and sharing microbial genomic data to foster interoperability

Publishing Title: A Schema for Digitized Surface Swab Site Metadata in Open-Source DNA Sequence Databases

Author Block: **A. B. Snyder**[1], B. Feng[1], D. Daeschel[1], D. Dooley[2], E. Griffiths[2], M. Allard[3], Y. Chen[3]; [1]Cornell University, Ithaca, NY, [2]Simon Fraser University, Burnaby, BC, CANADA, [3]U.S. Food and Drug Administration, College Park, MD.

Abstract Body: Large, open-source DNA sequence databases have been generated, in part, through the collection of microbial pathogens from swabbing surfaces in built environments. Analyzing these data in aggregate through public health surveillance requires digitization of the complex, domain-specific metadata associated with swab site locations. However, the swab

site location information is currently collected in a single, free-text “isolation source” field promoting generation of poorly detailed descriptions with varying word order, granularity, and linguistic errors, making automation difficult and reducing machine-actionability. We assessed 1,498 free-text swab site descriptions generated during routine foodborne pathogen surveillance. The lexicon of free-text metadata was evaluated to determine the informational facets and quantity of unique terms used by data collectors. OBO foundry library ontologies were used to develop hierarchical vocabularies connected with logical relationships to describe swab site locations. Five informational facets were identified: structure and subpart (253 unique terms), material construction (25 terms), condition (10 terms), and the orientation of the swab location on the structure (46 terms). Term hierarchy facets were developed as were statements (called axioms) about how entities within these five domains were related. The schema developed through this study has been integrated into a publicly available pathogen metadata standard, facilitating ongoing surveillance and investigations. The One Health Enteric Package is available at NCBI BioSample beginning June 2022. Collective use of metadata standards increases the interoperability of DNA sequence databases, enabling large-scale approaches to data sharing, artificial intelligence, and big-data approaches to food safety.

---PAGEBREAK---

Control Number: 2022-A-192-NGS

Topic 1: Epidemiological Cues: NGS in Clinical and Public Health Microbiology

Publishing Title: Utilizing the Terra Platform to Broaden Public Health Adoption of Best Practices in Pathogen Genomics

Author Block: **K. Libuit**[1], D. Park[2], J. Sevinsky[1];
[1]Theiagen Genomics, Highlands Ranch, CO, [2]Broad Institute, Cambridge, MA.

Abstract Body:
Public health pathogen genomics is a dynamic field with quickly evolving best practices. Access to bioinformatics solutions that capture these practices, however, is often limited to select institutions with advanced technical capabilities, familiarity with command-line environments, and robust compute resources. To broaden the adoption of our field’s best practices throughout the public health community, open-access resources that connect pipeline developers with pathogen-specific domain expertise to a user base with limited bioinformatics or programming experience is critical.

The Terra platform helps to address this challenge by providing a web-based, graphic user interface to containerized workflows. This interoperability has been exploited to establish a model for rapid development and continuous deployment of bioinformatics pipelines to end users in the public health community.

This model includes 1) the development of containerized workflows that capture best practices for analyzing critical pathogen, 2) provisioning of cloud compute resources to support Terra workspaces, 3) training of public health scientists on how to access and utilize these workflows through the Terra platform and 4) facilitation of regular communication between workflow developers and Terra users to enable continuous workflow improvements optimized for public health applications.

Throughout the COVID-19 pandemic, this model has been employed to establish SARS-CoV-2 bioinformatics capabilities in over 50 state and local public health laboratories (PHLs) across the US and dozens of PHLs and academic labs across West Africa and the Asia Pacific region. This work has directly supported hundreds-of-thousands of SARS-CoV-2 genome

submissions to internationally accessible databases and an ability to perform routine phylogenetic assessments to inform outbreak investigations in laboratories that previously lacked access to any bioinformatics resources.

Use of this model has also enabled the proliferation of the US Center for Disease Control and Prevention's MycoSNP workflow to public health users within two days of the initial request.

The Terra platform has helped to establish a model of bioinformatics development and distribution that bridges the gap between the technical microbial bioinformatics community and our front-line public health workforce. This approach is now being adopted in collaborative initiatives between workflow developers and various public health institutions to develop and distribute bioinformatics solutions for other pathogens of public health concern, including Enterics, healthcare-associated infections, multidrug resistant organisms, and Monkeypox.

---PAGEBREAK---