2024 ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines
October 13-16, 2024 • Washington, D.C.

NGS 2024 Oral Abstract Presentation Content and Schedule

Monday, October 14

9:15 AM – 10:00 AM | Scientific Session 1: Epidemiological Cues: NGS in Clinical and Public Health Microbiology

10:15 AM – 11:45 AM | Scientific Session 1: Continued: Epidemiological Cues: Oral Abstract Presentations

2:00 PM – 2:45 PM | Scientific Session 2: Bridging silos: Exploring mechanisms for collecting and sharing microbial genomic data for fostering interoperability

5:15 PM – 6:30 PM | Antimicrobial Resistance (AMR) Oral Abstract Presentations

---

Tuesday, October 15

9:15 AM – 10:00 AM | Scientific Session 3: Microbial Chatter: Microbial ecology in health and disease

10:15 AM – 11:45 AM | Scientific Session 3 Continued: Microbial Chatter: Oral Abstract Presentations

2:00 PM – 2:45 PM | Scientific Session 4: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

5:15 PM – 6:30 PM | Scientific Session #4 Continued: Secret Ingredient: Oral Abstract Presentations

---

Wednesday, October 16

9:15 AM – 10:00 AM | Scientific Session #5: Pipe Dreams: Analytical Methods, Bioinformatic Tools, and Pipelines

10:15 AM – 11:15 AM | Scientific Session #5 Continued: Pipe Dreams: Oral Abstract Presentations

---

Scientific Session 1: Epidemiological Cues: NGS in Clinical and Public Health Microbiology

- **Monday, October 14, 9:15 AM – 10:00 AM**

Oral Abstract Presentation #1:	**Phylogenomic analysis of colonizing and invasive Staphylococcus aureus in a neonatal intensive care unit reveals high levels of transmission of invasive strains**
Presenter:	**Qianxuan She**, MS, University of Pennsylvania/Children's Hospital of Philadelphia
Abstract Body:	Background: Staphylococcus aureus is a leading cause of healthcare-associated infection. Asymptomatic colonization by S. aureus represents a significant risk for subsequent infection. Recent data demonstrate that S. aureus continues to be one of the most common pathogens implicated in infections in the neonatal intensive care unit (NICU). Colonization and transmission of S. aureus amongst infants in the NICU could represent a significant risk of infection in this high-risk population. Methods: This study utilized whole-genome sequencing of 1,264 colonizing and invasive S. aureus isolates collected from infants hospitalized in the NICU at the Children's Hospital of Philadelphia over a two-year screening period. Comparative genomic analyses and genomic epidemiological approaches were implemented to investigate the transmission dynamics of S. aureus within the NICU setting. Results: Comparative genomic analyses identified a 73 single nucleotide polymorphism (SNP) threshold as a cutoff to demonstrate the close clonal population structure among NICU S. aureus genomes. Within this cutoff, 460 unique strains were identified, predominantly belonging to Clonal Complexes 8 and 398. 59 transmission clusters were inferred, with 16 invasive clusters involving 133 infants. The largest invasive cluster involved 24 infants, persisting over 2 years. Progression from colonizing to invasive clusters was observed, likely facilitated by an inability to eliminate colonizing strains. Spatial epidemiological links were found in some clusters, suggesting transmission among infants sharing NICU rooms, and some invasive clusters were potentially introduced through patient admissions. Compared to colonizing clusters, invasive clusters exhibited prolonged persistence and enhanced transmissibility, resulting in an increased number of patients involved within a specific time frame. Conclusions: These analyses have elucidated the genomic sameness among NICU S. aureus genomes and characterized the epidemiological links underlying the identified transmission clusters. Our findings demonstrate a clear association between colonization, transmission, and the development of invasive infections, underscoring the importance of implementing effective prevention and decolonization strategies to address S. aureus infections in the NICU setting.

Oral Abstract Presentation #2:	**Clinical and genomic characterization of Dengue virus during the 2023's outbreak in Mali, West Africa**
Presenter:	**Fousseyni Kane**, Ph.D., University Clinical Research Center (UCRC)-USTTB
Abstract Body:	Background: Dengue is a vector-borne disease prevalent in tropical regions, caused by the Dengue virus (DENV), which includes four serotypes: DENV-1 through DENV-4.

DENV infection has a range of outcomes from mild fever to severe complications, including hemorrhagic fever and shock syndrome. In Mali, following the 2019 DENV-2 serotype epidemic, a new outbreak began in August 2023. As of May 2024, Mali's Ministry of Health reports 9,976 suspected and 1,422 confirmed cases across 6 health regions, with 38 deaths, giving a 2.7% case-fatality rate among confirmed, symptomatic cases. We employed the pan-viral sequencing assay, VirCapSeq-VERT, at the University Clinical Research Center (UCRC) in Bamako, Mali, to characterize the current DENV outbreak. Methods: In accordance with national surveillance protocols, we collected samples from 23 patients with RT-PCR confirmed DENV, sequenced with VirCapSeq-VERT and bioinformatic analysis with the Rapid Identification of Microbes (RIM) pipeline. Bayesian analysis with Monte Carlo methods further explored the genomic sequences' temporal evolution. For vaccine and epitope analysis, the Dengvaxia vaccine strain was used in Geneious Prime 2023.1.2. Results: Among the 23 first cases of DENV identified during the current outbreak, 61% were male, with an average age of 43±13 years. The most prevalent symptoms included fever (83%), headache (57%), asthenia (48.5), anorexia (35%) and vomiting (30%). Additionally, several cases exhibited symptoms of gingival bleeding, anemia, jaundice and myalgia. Fifteen (62.2%) were identified as serotype 3 (DENV-3), followed by 4 (17.4%) serotype 1 (DENV-1). Four samples did not yield identifiable contigs. Time-calibrated phylogenetic trees revealed that the DENV-3 genomes identified in Mali share a common ancestor with DENV from Senegal collected in 2009, emphasizing the need to improve genomic surveillance of arboviruses in West Africa. Notably, divergent amino acids were observed throughout the prM and E proteins between the vaccine (Dengvaxia) and the wild-type strains in Mali, including 17 substitutions for DENV-1 and 14 for DENV-3. Conclusion: This study of Dengue virus (DENV) in Mali offers insights into molecular epidemiology and discrepancies between wild and vaccine strains. Most importantly, it underscores the imperative to enhance genomic surveillance of arboviruses in West Africa.

Scientific Session 1: Continued: Epidemiological Cues: Oral Abstract Presentations

- **Monday, October 14, 10:15 AM – 11:45 AM**

Oral Abstract Presentation #1: **Validation of an automated, end-to-end metagenomic sequencing assay for agnostic detection of respiratory viruses**

Presenter: **Amee Manges**, Ph.D., University of British Columbia

Abstract Body: Background: Current targeted molecular diagnostics, while cost-effective and clinically-actionable, are limited in the number and type of pathogens they can detect. Metagenomic next generation sequencing (mNGS) is an emerging approach to agnostic pathogen detection. Previous research has demonstrated the potential of mNGS for respiratory viral diagnostics. Translational barriers prohibit the widespread adoption of this technology in clinical and public health laboratories. Here, we validate an end-to-end mNGS assay for detection of respiratory viruses. Our assay is optimized to reduce turnaround time, lower cost-per-sample, increase throughput, and deploy secure and actionable bioinformatic results. Methods: Our assay is a modified version of a previously-published technique called Rapid-SMART9N. We optimized the protocol for sensitivity, throughput, and turnaround time. Our assay was automated using a Hamilton NIMBUS96 liquid handling robot. Clinical validation of the assay was performed using residual nasopharyngeal swabs submitted for routine diagnostic testing at Vancouver General Hospital and the BC Centre for

Disease Control (n = 359) that were RT-PCR-positive, or negative for Influenza, SARS-CoV-2, and RSV. Specimens were sequenced on a Nanopore GridION device for 3-hours, with subsequent data analysis on the BugSeq platform. We quantified sample stability, assay precision, the effect of background nucleic acid levels, and analytical limits of detection. Diagnostic performance metrics were estimated. Results: Assay optimizations resulted in a turnaround time of <12-hours for up to 55 multiplexed specimens. Significant differences in viral read counts across different days, operators, and reagent lots were not observed (Kruskal-Wallis: p = 0.66). Analytical limits of detection were estimated to be between 103-104 copies/mL for SARS-CoV-2, RSV, and Influenza. Our assay is highly specific (100%) and sensitive (61.9% overall and 86.8% for samples with RT-PCR Ct < 30). We reported incidental detection (4%) of viral pathogens in specimens, for which testing was not originally performed. Strain typing was assigned for 79% of mNGS positive specimens. Conclusions: We report the diagnostic performance of a rapid, sensitive, and high-throughput mNGS assay with automated data analysis for respiratory virus detection. This assay addresses translational barriers and represents an important step towards widespread adoption of mNGS for diagnostic, infection control, and public health surveillance uses.

Oral Abstract Presentation #2: **Genomic Epidemiology of Candida auris in the United States, 2013-2022**

Presenter: **Lindsay Parnell**, Ph.D., Centers for Disease Control and Prevention

Abstract Body:

Background: The pathogenic yeast Candida auris has become an urgent public health threat. Its unique ability to colonize skin and persist on surfaces for long periods of time has facilitated its spread within healthcare environments. Moreover, limited treatment options due to frequent multi-drug resistance is cause for serious concern. In the public health setting, genomic sequencing has been utilized to monitor circulating clades and strains, supplement outbreak investigations, and identify resistance mechanisms. To date, six clades have been reported globally. A national picture, providing updated genomic epidemiologic trends in the United States, has not been reported since 2017.Methods: We obtained and performed whole genome sequencing on a convenience sample of 1,540 C. auris isolates, primarily collected through the Antimicrobial Resistance Laboratory Network. Isolates included in the dataset were both geographically and temporally diverse — representing cases from all seven U.S. regions collected between 2013 and 2022.Quality control and single nucleotide polymorphisms (SNP) analyses were performed using MycoSNP (v1.4), a reference-based bioinformatics pipeline for fungal organisms. Genomic results were integrated with detailed epidemiologic metadata to describe phylogeographic and phylotemporal trends and clade-specific profiles related to antifungal resistance and genetic diversity. Preliminary Results: Isolates in this U.S. study belonged to clades I (56%), II (<1%), III (29%), and IV (14%). Using clade-specific references, we found that pairwise SNP distances ranged from 0 to 1004 in Clade I, 11 to 2323 in Clade II, 0 to 105 in Clade III, and 0 to 250 in Clade IV. Within each clade, we observed isolates clustering by geographic region of collection. In some instances, we observed high relatedness between isolates collected in different geographic regions. C. auris resistance-related mechanisms were identified, including ERG11 Y312 and FKS1 S639P.Conclusions: Intra- and inter-regional relatedness reinforce the role of both local spread and travel-related importation as important contributors of transmission in the United States. We show the absence of Clade V and the recently reported Clade VI. We

observed azole and echinocandin-resistance mechanisms across multiple clades. Continued genomic surveillance of C. auris will improve our understanding of transmission dynamics and strengthen preparedness and response to this emerging fungal threat.

Oral Abstract Presentation #3: **Enrichment Sequencing in Aircraft Wastewater Samples Efficiently Captures Norovirus Global Circulating Genetic Diversity**

Presenter: **Xueting Qiu,** Ph.D., Ginkgo Bioworks

Abstract Body: Background: Norovirus causes an estimated 20% of acute gastroenteritis cases annually. Despite extensive testing networks, global genomic surveillance of Norovirus is minimal. Aircraft wastewater surveillance offers rapid detection, characterization, reporting, and forecasting of important pathogens from diverse origins without needing to establish surveillance in every country. We evaluated aircraft wastewater data for its utility in capturing the global circulating genetic diversity of Norovirus and filling existing surveillance gaps. Methods: Aircraft wastewater was collected daily from 25 flights, 6 days/week, at an airport in the Middle East from November 2022 through April 2024, resulting in >5000 samples representing 70+ countries. A subset of samples (n=435) spanning a 13-week timeframe during the epidemic season of 2023-2024 was sequenced using a >60 targeted viral surveillance enrichment panel, including Norovirus genogroups GI, GII, GIV, and Norwalk-like viruses. Bioinformatic analysis was conducted using custom end-to-end pipelines (QA/QC, variant calling, lineage assessment) to characterize and contextualize results. Results: The 435 samples represented a broad geographic distribution, encompassing 61 distinct countries. The samples yielded genomes for 16 pathogen groups, including SARS-CoV-2, Influenza, RSV, Adenovirus, and Norovirus. Over the 13 weeks in 2023-2024, enrichment sequencing yielded a total of 30 Norovirus whole genomes capturing the diversity of circulating strains of genogroup GI (n = 1), GII (n=22), and GIV (n=7). Whole genomes have not been deposited in NCBI for several seasons for GI and GIV. Phylogenetic analysis demonstrates that GII sequences from aircraft wastewater clustered closely with public sequences collected from the 2023-2024 season but represent a diversity of circulating strains. Public sequences of GII had very biased geographic representation with only 24 genomes from 3 countries (Japan, Thailand, and Canada). Here, we demonstrate the recovery and characterization of 22 GII genomes from aircraft wastewater representing 13 countries of origin in only 13 weeks of surveillance. Conclusions: Aircraft wastewater surveillance can effectively fill global genomic surveillance gaps for Norovirus. The genetic diversity captured by this study serves as a valuable resource that can be applied to intervention decisions, vaccine design, and therapeutic development.

Oral Abstract Presentation #4: **Direct-From-Specimen Nanopore Sequencing for Rapid Detection of TB Drug Resistance**

Presenter: **Jennifer Guthrie**, Ph.D., University of Western Ontario

Abstract Body: Background: Tuberculosis (TB) is a leading cause of death worldwide, particularly impacting low-to-middle income countries. Drug resistance in Mycobacterium

tuberculosis (Mtb), the agent responsible for TB, poses significant challenges, particularly in regions with high prevalence. The slow growth rate of mycobacteria means that phenotypic susceptibility testing can take weeks to months. However, comprehensive genomic and bioinformatics approaches, combined with direct-from-specimen methods, offer promising avenues for rapid and accurate resistance detection. The Oxford Nanopore Technologies MinION is key in this effort, providing a portable sequencing solution for remote and rural areas with limited laboratory infrastructure, which are highly impacted by TB. Methods: In pursuit of improved methodologies for rapid detection of Mtb resistance we used a lysis buffer containing a patent-pending mycobacteria-specific enzyme to extract DNA directly from sputum specimens. We then designed an amplicon-based protocol to sequence key regions of the genome containing resistance conferring mutations, which was used with the DNA extracted directly from sputum specimens. Next generation sequencing (NGS) of the amplicons was conducted using a MinION. A bioinformatics pipeline was developed for data analysis, which includes quality control measures and identification of resistance mutations with a minimum sequencing depth of 10X. The results are compiled into a comprehensive report for further analysis. Results: The lysis enzyme method resulted in 15 ng/ul (SD 0.42) of mycobacteria genomic DNA, yielding a single contig by NGS compared to 6 ng/ul (SD 0.1) and over 80 contigs by standard bead beating methods. Primers were designed to amplify 17 genomic regions known to harbour resistance conferring mutations, achieved across 21 amplicons using multiplex PCR. By optimizing primer concentrations and combinations, all targets were amplified using three multiplex PCR reactions, which were pooled for NGS. All target regions were sequenced with a mean depth of >200X and 100% coverage, allowing for the identification of resistance mutations. The bioinformatics pipeline successfully processed the sequences, outputting results as a text file for easy database import and a comprehensive PDF report. Conclusions: These results bring us closer to rapid Mtb antimicrobial resistance detection, with significant implications for the diagnosis and treatment of resistant infections and enabling a shorter time to effective therapies.

Oral Abstract Presentation #5: **Implementation of regional sharing of WGS data for detection of nosocomial bacterial clusters**

Presenter: **Lieke van Alphen**, Ph.D., Maastricht University Medical Center

Abstract Body:

Background: The identification of clonal clusters of pathogens is essential to support the epidemiological investigation of nosocomial outbreaks, which is typically conducted independently by hospitals. However, patient transfer within a region putatively spreads antimicrobial resistant (AMR) isolates among institutions. Real time data sharing on (potential) outbreak isolates is necessary to gain a better understanding of the AMR pathogen transmission between hospitals. We set up regional networks for data sharing of whole genome sequence (WGS) data in two distinct regions of the Netherlands, allowing rapid cluster analysis. Methods Organisms of choice were determined per region and participating hospitals (Region 1: 4 hospitals, Region 2: 2 hospitals) sent in isolates with limited metadata. Protocols were established for transfer of biological material, routine processing of isolates and real-time sharing of data. WGS of nosocomial bacterial isolates was performed using Illumina technology in central typing facilities. Data was uploaded in the cloud-based platform of 1928 Diagnostics (separate databases per region),

where each participating institution could access the shared data. In this platform, identification of AMR genes and analysis of intra- and inter-hospital clusters using cgMLST were performed. Where necessary additional SNP analysis was performed. Results: Service-level agreements were established in each region detailing the conditions and obligations for the institutions participating in the network. Vancomycin-resistant Enterococcus faecium and ESBL-Klebsiella pneumonia isolates were included in both regions. Region 1 additionally included other ESBL-positive bacteria. Region 1 analysed 845 isolates in 2,5 years and region 2 334 isolates in 1 year. The data sharing platform proved user-friendly for regional partners, enabling rapid access to AMR gene and cgMLST data of the hospital isolates and analysis of local isolates in a regional context. It also facilitated WGS implementation in non-academic regional hospitals without dedicated bioinformaticians. Conclusion: Regional data sharing of WGS data was feasible whilst AMR gene information and cluster analysis was rapidly available for each regional hospital, although the turn-around time of WGS should be optimised. Phylogenetic analysis of the results provided for the first time an overview of the population diversity of nosocomial isolates at the regional level. Implementation of this project resulted in regional WGS capacity building

Scientific Session 2: Bridging silos: Exploring mechanisms for collecting and sharing microbial genomic data for fostering interoperability

- **Monday, October 14, 2:00 PM – 2:45 PM**

Oral Abstract Presentation #1: **A Novel Format for Capturing Cell Organization**

Presenter: **Nabil-Fareed Alikhan**, Ph.D., University of Oxford

Abstract Body: Background: Understanding the genomic dynamics and content of bacteria, fungi, and other cellular life is crucial for deciphering patterns of evolution, adaptation, and pathogenicity. However, in bacteria specifically, the diversity and complexity of mobile genetic elements (MGEs) present challenges in data representation and analysis. Methods: We present a novel human and computer-readable format designed to comprehensively describe the genetic content within bacterial cells. The format can encompass any combination of chromosomal and accessory genomic content including plasmids, transposons, integrons, and bacteriophages, along with associated metadata such as genetic annotations, host information, and mobility characteristics. The proposed format employs structured data elements organised hierarchically, facilitating efficient storage, retrieval, and analysis of MGE-related information. Results: To demonstrate the utility of this format, we provide a case study involving the annotation and comparison of MGE content from a set of simulated and Klebsiella pneumoniae genomes. Through this case study, we illustrate the effectiveness of the format in capturing intricate MGE arrangements and identifying potential horizontal gene transfer events and evolutionary relationships among bacterial strains. Conclusions: Overall, the proposed Wolvercote format offers a systematic approach to describe and analyse MGE content in bacterial cells, fostering advancements in microbial genomics, evolutionary biology, and genomic epidemiology. Its adoption can streamline data management, enhance reproducibility, and accelerate discoveries pertaining to the role of MGEs.

Oral Abstract Presentation #2: **Pathoplxus and Loculus: A New Platform for Sharing Pathogen Sequencing Data**

Presenter: **Chaoran Chen**, Ph.D., ETH Zurich

Abstract Body:

This platform is not yet publicly available but will launch soon. We would appreciate it if the abstract committee could keep our topic confidential until the end of June.Pathogen sequencing data are crucial for understanding epidemics, for providing insights to pathogen evolution, as well as for the development of vaccines and treatments. These data are a crucial resource for public health and research. Unfortunately, there are several hurdles to rapid sharing on a global scale. It has been challenging to ensure that data generators are properly attributed for their work and retain their opportunity to be the first to publish with their own data (avoiding "scooping") while making the data available to facilitate effective downstream analyses for public health benefits and ensure long-term perseverance of the data. Furthermore, submitting, browsing, and retrieving data of large datasets - is a major challenge. Our new platform, Pathoplexus, addresses these issues.Pathoplexus provides a flexible data-sharing scheme where data providers can choose to share their data openly or with time-limited protections. It integrates with existing INSDC databases by enabling Pathoplexus 'open data' to synchronize with INSDC, and INSDC data to be accessed through Pathoplexus. Pathoplexus initially prioritizes viruses for which many sequences are not being shared or shared through general platforms such as GitHub and a dedicated platform has been missing. It will be launched for the Ebola, the West Nile and the CCHF virus and may be expanded in the future.Pathoplexus is powered by a newly developed software called Loculus. Loculus is an open-source, general-purpose software package for sharing, maintaining, and accessing microbial genetic sequencing data. Loculus can be used by any group managing sequencing data, from the level of individual public health or research laboratories right up to international databases facilitating global pathogen sequence sharing.Pathoplexus was created by a working group of PHA4GE (Public Health Alliance for Genomic Epidemiology), an international group working to establish better standards for public health and bioinformatics, and additionally, has an independent executive board that is dedicated to running Pathoplexus according to its charter. In addition, Pathoplexus has a scientific advisory board providing advice to the executive board. Pathoplexus' charter and guidelines for the boards are publicly available and prioritize transparency and fairness.

Oral Abstract Presentation #3: **Contextual Data Reporting in Infectious Disease Genomic Epidemiology for Outbreak Investigation and Surveillance: A Scoping Review**

Presenter: **Tian Rabbani**, MPH, Simon Fraser University

Background: Practitioners interested in infectious diseases have traditionally used microbiological cultures or molecular biomarkers to discern the transmission of pathogens among a population. Next-generation sequencing (NGS) technologies have allowed for novel infectious pathogens to be tracked at a molecular level effortlessly. However, sequence data alone seldom reveal a pathogen's virulence and transmission in the absence of accompanying metadata from the host and sample. This review summarizes the current landscape of contextual data in studies that use NGS for outbreak or surveillance investigation. Methods: References were identified using PubMed and Embase and restricted to year 2000 and forward, English language, and human species. In sum, 4,031 articles were identified of which 3,386 were screened after de-duplication and 621 were screened for full-text eligibility. A total of 499 articles were included and manually assessed for the availability of 87 metadata types. Results: Included studies focused on surveillance (n=230, 46.1%), outbreaks (n=223, 44.7%), or both (n=46, 9.2%) with a representative diversity of 64 genera. Genome sequence data were not publicly accessible for all studies (n=103, 20.6%). The availability of five overarching non-sequence derived metadata fields, and granular subfields, were investigated in-text, repositories, and supplementary files. Antimicrobial resistance (n=239, 47.9%) and gene (n=313, 62.7%) metadata were least reported. Sample collection metadata was reported by all studies, with host and sequencing methods reported by 98.0% (n=489) of studies. Among the mandatory attributes outlined by NCBI's Pathogen BioSample Package, sample geographical location (n=486, 97.4%), country (n=484, 97.0%), and collection date (n=430, 86.2%) were the most reported. Moreover, host scientific name (n=268, 53.7%), sample collection by (n=285, 57.1%), and host disease (n=322, 64.5%) were the least reported attributes. Conclusions: Consistent metadata reporting improves (re-)use of both contextual and sequence-derived data. Evidently, barriers pose when samples are uploaded with missing mandatory attributes. Temporal data, such as time and place, are pivotal to robust public health response however, additional metadata allows for a complete picture of transmission dynamics, which were not consistently reported in the literature. This review highlights gaps in metadata reporting in genomic epidemiology both when submitting to peer-reviewed journals and uploading to repositories.

Antimicrobial Resistance (AMR) Oral Abstract Presentations

- **Monday, October 14, 5:15 PM – 6:30 PM**

Oral Abstract Presentation #1: **Population Diversity of Matched Gastrointestinal and Bloodstream Populations from Patients with Vancomycin-Resistant Enterococcus faecium Infection**

Presenter: **Emma Mills**, University of Pittsburgh

Abstract Body: Vancomycin-resistant Enterococcus faecium (VREfm) is a prevalent hospital-acquired pathogen, primarily transmitted via the fecal-oral route. VREfm overgrowth in the gastrointestinal (GI) tract can lead to difficult-to-treat bloodstream infections (BSI) due to translocation into the bloodstream. To characterize the diversity and identify adaptive traits between these environments, we performed population-level sequencing of matched (≤ 7 days apart) GI and BSI VREfm samples, pooling 100-1000 colony forming units per VRE-selective plate, from 16 patients with BSI. We found all 16 patients had VREfm in their BSI populations, with one patient showing a

GI tract co-colonized with VREfm (70%) and vancomycin-resistant E. faecalis (30%). To identify within patient multi-locus sequence type (ST) diversity, we compared the population sequences of GI and BSI populations to patient-specific reference genomes from a single clone from the BSI population. In 11 of 16 patients (69%), BSI and GI populations showed singular and matching STs, suggesting the BSI infection likely originated from the GI reservoir. In the 5 patients with ST discrepancies, 2 patients showed BSI and GI populations of singular but different STs, suggesting the BSI may have resulted from an external source. In the remaining 3 patients, we identified a mixture of two STs within the GI population, one of which was also present in the BSI population. In the 11 patients with singular and matching BSI and GI populations, we assessed the diversity of each population by identifying variants present at 5-100% frequency compared with the patient-specific reference genome. We found a numerically higher though nonsignificant diversity in the GI populations, with an average of 17.1 variants/population compared to BSI population (13.6). BSI populations exhibited statistically significant enrichment of mutations in amino-sugar metabolism genes (Plt; 0.00002, Bonferroni adjusted), with variations in the nanE gene found in 3 patients. Notably, mutations in daptomycin-resistant genes clsA (n = 2) and mprF (n = 1) were identified in 3 patients, only in their GI populations (P =0.004). In both GI and BSI populations, DNA repair pathway gene umuC was identified across 3 patients suggesting an advantage in both environments (P& = 0.0001). Together, this study demonstrates the utility of population-level sequencing to characterize the genetic diversity and identify potential adaptive traits in both GI and BSI environments within patients with VREfm BSI.

Oral Abstract Presentation #2: **Dogs Fed Raw Meat-Based Diets as Vectors of Extensively Drug-Resistant *Salmonella4*,[5],12:i:- From Cattle to Infants**

Presenter: **Isabelle Bernaquez**, MS, Laboratoire de santé publique du Québec

Abstract Body:

Background: Salmonella enterica serovar 4,[5],12:i:- (S. 4,[5],12:i:-), a monophasic variant of S. Typhimurium, recently became the 3rd most frequent serovar in Canada of nontyphoidal Salmonella (NTS). Extensive drug-resistance (XDR), defined as resistance to ampicillin, ceftriaxone, azithromycin, ciprofloxacin and trimethoprim/sulfamethoxazole, is still rarely detected among NTS. Here, we describe an in-depth investigation conducted in Quebec, Canada as part of the response to a multi-provincial XDR S. 4,[5],12:i:- cluster. Methods: Cluster detection was done via whole genome sequencing and phylogenetic analysis for PulseNet Canada. Human cases were investigated with a standardized epidemiologic questionnaire. Subsequent animal and targeted food inspection investigations were carried out. Antimicrobial resistance genes (ARGs) and plasmids were predicted by Staramr and MOB-suite, and confirmed by antimicrobial susceptibility testing and long-read sequencing. Global phylogenetic comparisons were performed in NCBI Pathogen Detection. Results: From July 2020 to June 2023, we identified 20 human cases, 16 veal calves, 3 dogs, one piglet, one moose, and 4 raw meat-based diet (RMBD) isolates in Quebec. Infants (≤ 2 years) were mainly affected (50%), and dogs fed a RMBD was the major source identified, followed by exposure to cattle. The emergence of new cases in July 2021 coincided with the acquisition of a conjugative XDR IncHI2A plasmid encoding 13 ARGs [including mph(A), blaCTX-M-55 and qnrS1] and possible reduced biocide and heavy metal susceptibility via efflux pump regulators (ramAp, marR), oxidative stress response (umuDC, dsbC), protective DNA-binding proteins (hns, hha) and metal resistance (terZABCDE, copG). Environmental persistence is thus likely due to co-selection pressures. Different

sources (e.g. animal feed) were detected for this plasmid compared to other IncHI2A plasmids in public databases (e.g. poultry), indicating different niches for different plasmid subtypes. Two other plasmids were also detected, which encoded putative phage defenses and kanamycin resistance. Global phylogenetic analysis found an association with cases and pork products from USA. Conclusions: We report the first occurrence of XDR S. 4,[5],12:i:- in humans associated with RMBDs for dogs. This study highlights the threat of XDR S. 4,[5],12:i:- ST34 to vulnerable populations, where impacted veal farms and asymptomatic dogs can act as disease carriers with limited treatment options: a One Health concern.

Oral Abstract Presentation #3: **Data Visualization Connecting Genomic Analysis, Test Result Data and Patient Demographics for Carbapenemase-Producing Organisms and Mycobacterium tuberculosis**

Presenter: **Kimberlee Musser**, Ph.D., Wadsworth Center, NYSDOH

Abstract Body: Background- Visualization of next generation sequencing (NGS) data in combination with other clinical test data and patient demographics is often challenging as NGS data are not typically integrated into laboratory information management systems (LIMS). Additionally, clinical LIMS often lack the capabilities to provide visualizations so trends can be tracked and viewed in real time allowing the best use of these data. Methods-We have performed whole genome sequencing (WGS) > 13,000 carbapenemase-producing organisms (CPOs) and Mycobacterium tuberculosis (MTB) and analyzed using in-house developed bioinformatics pipelines which confirm bacterial species, provide typing (e.g., multi-locus sequence type (MLST), MTB lineage and spoligotype), and identify predictors of antimicrobial resistance (AR; e.g., AR gene variants and mutations present). The pipelines are run on a weekly basis on all CPO and MTB isolates that undergo WGS, and outputs including pipeline version, sequencing read file names, quality control metrics, typing data and AR gene and mutations identified are uploaded to our LIMS. We are expanding our data visualization options to enhance our utilization of this WGS data and are working with the Digital Epidemiology Services and the Centre for Genomic Pathogen Surveillance (CGPS) to create dashboards for both CPOs and MTB using Data-flo and Microreact software, beneath our firewall. Results-These new dashboards are creating user-friendly interactive views of real-time trends in data outputs. The software tools transform, summarize and visualize these data to enhance data exploration and streamline interpretation. We have views that allow visualization over time, mapping, by bacterial species, MLST, lineage, spoligotype, AR gene or mutation, as well as specimen type, tests requested and other patient demographic and laboratory testing data. Conclusions-Visualization of outputs of NGS pipelines in combination with existing data from systems such as LIMS is providing efficient ways to summarize and view that data in real-time. With more than a dozen dashboards developed as well as collaboration with epidemiology partners, these projects demonstrate a path forward for comprehensive integration of NGS data into public health workflows. We strive to expand the utilization of NGS data through providing informative dashboards that can be shared regularly with partners and monitored for trends that may be missed and ultimately hope to demonstrate clinical and public health impact.

Oral Abstract Presentation #4: **Development of a Plasmid Reconstruction Pipeline and Application to Track Spread of Carbapenemase-Containing Plasmids in a Long-Term Acute Care Hospital**

Presenter: **Auden Bahr**, University of Michigan Medical School

Abstract Body:

Background: An important mechanism for the spread of carbapenem resistance is plasmid-mediated horizontal gene transfer. While evidence for horizontal gene transfer has been observed in hospitals, current challenges associated with large-scale studies of healthcare-associated plasmid transfer include accurately assembling plasmid sequences and identifying recent plasmid transfer events between patients. Methods: Rectal surveillance swabs were conducted upon admission and every 2 weeks on patients admitted to a Chicago-area long-term acute care hospital between June 2012 to June 2013 with 94% adherence. In total, 551 isolates from 261 patients were subjected to Illumina short-read sequencing, with a subset of 36 isolates additionally subject to Nanopore long-read sequencing. MOB-suite was used to reconstruct plasmids from short-read sequencing data and evaluated using hybrid assemblies. A method to identify a representative set of isolates for long-read sequencing to enable tracking of plasmids was developed using MOB-suite and Mashtree. Results: The majority (83%) of isolates were Klebsiella pneumoniae, and 93% of these were predicted by AMRFinder to carry blaKPC located on a plasmid. Using MOB-suite's standard plasmid reference databases led to poor plasmid reconstruction with short-read data. Only 60% of short-read plasmid bins carried the correct assortment of antibiotic resistance genes, and 52% contained sequences from multiple plasmids or the chromosome. An additional subset of 77 isolates was identified with the developed approach that captures the genetic diversity of the plasmids, and hybrid assemblies were constructed to supplement the reference database with the aim of improving reconstruction accuracy. Conclusions: Short-read data can be used to identify gaps in plasmid databases, as well as inform selection of isolates for long-read sequencing to enable plasmid tracking and improve accuracy of plasmid reconstruction. Work is ongoing to leverage the dense sampling of the facility to understand the dynamics of plasmid transfer and spread in a high-prevalence setting. Results will yield data generation and analysis strategies to empower more effective studies of genomic epidemiology of plasmids in healthcare settings.

Scientific Session 3: Microbial Chatter: Microbial ecology in health and disease

- **Tuesday, October 15, 9:15 AM – 10:00 AM**

Oral Abstract Presentation #1: **The positive impact of a CSF metagenomic next-generation sequencing assay on diagnosis and management of neonatal Metamycoplasma hominis meningitis**

Presenter: **Benjamin Liu**, Ph.D., Children's National Hospital, George Washington University

Abstract Body:

Background: Metagenomic next-generation sequencing (mNGS) for pan-pathogen detection in cerebrospinal fluid (CSF) has shown promise in aiding diagnosis of central nervous system (CNS) infections. However, the clinical impact of a CSF

mNGS test on the diagnosis/management of CNS infections in children is difficult to ascertain due to the lack of orthogonal microbiology confirmation and complexity for many of these cases. We identified Metamycoplasma hominis by CSF mNGS test in an 11-day old male born at 383/7 weeks who presented with fever, intermittent posturing and acute/subacute intraparenchymal hematoma. No infectious etiology had been determined in prior CSF studies. Due to the patient's close contact with tuberculosis, empiric TB treatment was started 7 days before the CSF collection. We performed further molecular and microbiology testing and chart review to determine the clinical impact of the M. hominis positive mNGS on the subsequent management of this patient. Methods: The CSF positive for M. hominis by mNGS was subjected to M. hominis PCR and Mycoplasma culture using 10B broth and A8 agar. Antimicrobial susceptibility testing (AST) was performed for a M. hominis isolate from the CSF culture. Longitudinal follow-up of CSF cell counts, chemistry indices, and the patient's clinical status and neurologic condition were monitored during and after treatment. Results: The CSF mNGS assay detected ~600 reads covering 7% of the M. hominis genome. As a result, empiric TB treatment and a planned brain biopsy were cancelled and targeted therapy with levofloxacin was continued. Due to the limited mNGS coverage, another aliquot of CSF was tested by Mycoplasma culture and M. hominis PCR, both of which were positive. M. hominis AST showed susceptibility to tetracycline, moxifloxacin, levofloxacin, and clindamycin. CSF cell counts and chemistry indices, including glucose, nucleated cells and proteins returned to normal after 4 weeks of clinical follow-up. The patient's clinical status and neurological conditions were also markedly improved. Conclusions: The identification of M. hominis by mNGS was confirmed by subsequent molecular and microbiology testing, including culture and AST. Clinical follow-up demonstrated positive impact of the mNGS assay on the diagnosis and management of the neonatal M. hominis meningitis case by avoidance of an invasive brain biopsy and enabling the administration of targeted and effective treatment for the patient's CNS infection.

Oral Abstract Presentation #2: **The Molecular Space Age: Unlocking the Microbiome of the International Space Station within Sito Nanopore Sequencing**

Presenter: **Sarah Castro-Wallace**, Ph.D., NASA

Abstract Body:

Background: Since 2000, the microbiome of the International Space Station (ISS) has been monitored to assess risk to both crewmembers and the spacecraft through onboard culture-based methods with ground-based analyses. While this approach has served to provide alerts to anomalies and increase confidence in operational controls, these data are biased toward microorganisms that grow on the singular media type and ISS-available growth conditions; this leads to the false depiction of an overall lack in biodiversity. As NASA moves to a focus on exploration beyond low-Earth orbit where routine sample return will not be possible, it is critical to understand the microbiome of the spaceflight environment and its possible association with the noted positive influence on crew and vehicle health by using in situ methods for routine monitoring. Methods: A culture-independent, nanopore sequencing-based method has been implemented onboard the ISS and for returned swabs. DNA is directly extracted from the swab tip, purified, amplified, amplicons purified, libraries prepared, and nanopore sequencing is initiated. A detailed statistical comparison of the more than 20 years of culture-based data to the culture-independent nanopore data was performed. Results: Data from both methods similarly describe a human-occupied environment. Both data sets depict a

common core microbiome across time and location, but the nanopore data describe an expanded microbiome with increased diversity of human- and environmentally-associated microbes. Additionally, through detailed analysis of the nanopore-generated data, a few ISS locations are emerging as unique ecological niches, potentially resulting from environmentally-driven microbial selection. The presence of some noted taxa within these unique locations has implications for crew health, planetary protection, and controls used in future spacecraft systems. In terms of data usability, an evaluation of the NASA risk assessment process found that, regardless of the data set utilized, the risk assessment was identical. Therefore, use of the nanopore data was not found to drive more frequent remediation or alter the number of remediation events. Conclusions: The ability to perform in situ molecular microbial profiling is transforming how NASA assesses risk and is a critical tool towards monitoring and maintaining the microbiome of exploration spacecraft.

Scientific Session 3 Continued: Microbial Chatter: Oral Abstract Presentations

- **Tuesday, October 15, 10:15 AM – 11:45 AM**

Oral Abstract Presentation #1: **Persistent Salmonella Infections in Humans are Associated with Mutations in the Bara/Sira Regulatory Pathway**

Presenter: **Alexandra Grote**, Ph.D., The Broad Institute of MIT and Harvard

Abstract Body:

Background: Many clinically important bacterial pathogens, including Salmonella enterica, are able to establish persistent infections in humans. These persistent bacteria can evade the host immune system and killing by antibiotics, and provide a reservoir of disease for infection of healthy individuals. Non-typhoidal serovars of S. enterica are estimated to cause 93.8 million cases of gastroenteritis each year, resulting in approximately 155,000 deaths and can establish and maintain symptomatic and asymptomatic chronic human infections that have been linked to inflammation and cancer. Despite the significant global health burden associated with S. enterica infections, the underlying mechanisms of Salmonella persistence are poorly understood. Methods: From the largest retrospective collection of Salmonella to date, we whole genome sequenced 658 S. enterica isolates obtained from patients in Israel presenting with persistent salmonellosis between 1995 and 2012. Each of the 256 patients was represented by 2-5 longitudinal isolates over the course of infections that lasted between 30 and 2,001 days from the first culture-confirmed diagnosis. We determined mutations between each patient's earliest isolate and each subsequent longitudinal isolate. Results: We found a significant enrichment of mutations arising in global regulatory genes, including mutations in the BarA/SirA two-component regulatory system in 24 different patients. Comparative RNA-Seq analysis revealed that distinct mutations in barA/sirA led to diminished expression of Salmonella Pathogenicity Islands 1 and 4 genes, which are required for Salmonella invasion and enteritis. Moreover, competition experiments in the acute salmonellosis mouse model revealed that the late isolates harboring a mutation in either barA or sirA were significantly less virulent than the early isolates. Using RNA-Seq on infected mouse BMDMs, we found that the transcriptional host response to these mutants was defined by a significant downregulation of genes involved in the immune, defense, and cytokine responses. In the persistence mouse

model, we found that at later time points, barA/sirA mutants were colonizing intestinal and systemic sites and being shed at comparable or even higher levels than the wild-type strain. Conclusions: Taken together, these findings suggest that selection of mutations in global virulence regulators facilitates persistent Salmonella infection in humans, by attenuating Salmonella virulence and inducing a weaker host inflammatory response.

Oral Abstract Presentation #2: **Gut Microbiome of the Orang Asli in Malaysia: Expanding Global Diversity**

Presenter: **Soo Ching Lee**, Ph.D., National Institute of Allergy and Infectious Diseases

Abstract Body:

Background: With advancements in sequencing and computational approaches, recovering metagenome assembled genomes (MAGs) is an important tool in microbiome studies to provide more comprehensive reference databases. However, microbiome catalogues are still under-represented for indigenous communities worldwide, especially those in Southeast Asia. Methods: We utilized shotgun metagenomic sequencing to reconstruct prokaryotic and DNA viral MAGs from a total of 650 fecal samples (~20M reads/sample), collected longitudinally from 351 indigenous (Orang Asli) and 56 urban Malaysian subjects. Additionally, we performed metatranscriptomic sequencing on a subset of 96 indigenous and 30 urban fecal samples (~40M reads/sample) to reconstruct the RNA viral MAGs. All reads were assembled with metaSPAdes, followed by MEGAHIT for unassembled reads. These combined assemblies were binned using MetaBAT2, MaxBin2, and CONCOCT to generate the prokaryotic MAGs. We used Cenote2 for viral identification. Results: For prokaryotic MAGs, we integrated the assembled genomes with Unified Human Gastrointestinal Genome (UHGG) and Korean, Indian, Japanese (KIJ) database MAGs to generate an expanded catalog, which contains 5,355 representative prokaryotic species, including 307 (5.7%) unique to the Malaysian dataset, predominantly under the Clostridia class. Additionally, we identified 51,179 viral sequences, representing 13,829 viral operational taxonomic units (vOTUs) after dereplication of sequences with >=95% sequence identity. Taxonomic annotation on bacteriophages using vConTACT2 revealed that a significant portion of the vOTUs (96.1% of 13,829) were not aligned to any known genomes at the genus level. Almost half of the vOTUs (6,278; 45.5%) are unique to the Malaysian dataset compared to other recently published virome databases. Furthermore, we detected gut bacteriophages that more abundant in helminth infected Orang Asli (OA). Metatranscriptomic analysis on a subset of 96 fecal samples identified human RNA viruses (eg. Norovirus and Enterovirus) that were more prevalent in the OA compared to urban residents in Kuala Lumpur. Conclusions: By analyzing the microbiome of an under-represented indigenous population, we significantly expand our current databases representing gut microbial ecology, which is more representative of individuals living in urban industrialized societies. Our research highlights the value of including more under-represented populations globally to capture the full spectrum of human microbiome variation.

Oral Abstract Presentation #3: **Impact of a GABA-producing probiotic on the bacterial and fungal microbiota during CNS inflammatory demyelination**

Presenter: **Javier Ochoa-Repáraz**, Ph.D., Boise State University

Abstract Body:

Background: The gut microbial populations are exposed to significant changes during inflammatory demyelinating diseases and experimental models of diseases. In experimental autoimmune encephalomyelitis (EAE), the most used animal model to study multiple sclerosis (MS) changes in the abundance of bacterial taxa has been documented by us and others. Some of those changes affect bacteria capable of producing gamma-aminobutyric acid (GABA). Because of the importance of GABA as an inhibitory neurotransmitter and as an immunomodulator factor that regulates immune cell function and decreases inflammation, we genetically engineered a Lactococcus lactis strain to overproduces GABA. Methods: C57BL/6 mice EAE mice were treated with 5x109 colony-forming units (CFU) of GABA-producing L. lactis, unmodified L. lactis, or with sham, in combination or not of glutamic acid, the substrate for glutamic acid decarboxylase, the enzyme responsible for GABA production in L. lactis (all 6-7 mice/group). The treatments were performed daily by oral gavages. EAE was induced on day 0 (0 days post-induction: dpi), while treatments started seven days before and lasted for the duration of the experiment. Stool samples were collected at the beginning of the treatment (-7 dpi), EAE induction (0 dpi), 14 dpi, and 28 dpi. 16S rRNA (bacteria/archaea) and IST (fungi) amplicon sequencing were performed using all samples collected. In addition, PICRUSt2 analysis was done to predict functional abundances. Results: Our results indicate that the oral treatment with GABA-producing L. lactic resulted in significant alterations in the microbiota and mycobiota composition during EAE, with significant changes in the differential abundances of specific taxa. Conclusions: Our study first assesses the potential impact of a genetically engineered probiotic on both gut microbiota and mycobiota.

Oral Abstract Presentation #4: **Using Positive Selection-Augmented Next Generation Sequencing to Interrogate Wastewater for Viruses and Pathogenic Bacteria Circulating in a Population**

Presenter: **John Collins**, MPH, Columbia University Mailman School of Public Health

Abstract Body:

Background Wastewater surveillance is a noninvasive way to assess the prevalence of pathogens within a community. Circulating viruses from infected humans and animals are collected into local wastewater systems, so they serve as an ideal source for interrogation using efficient next generation sequencing (NGS) systems providing real-time data for public health monitoring. Innovative pan-pathogen metagenomics detection tools developed have demonstrated the power of positive selection capture probe-based systems targeting the entire genome of all vertebrate viruses (VirCapSeq-VERT) and the core genome of pathogenic bacteria/AMR elements (BacCapSeq). Capture-sequencing is known to increase signal-to-noise by up to 1000-fold, driving down cost and easing the bioinformatical analyses of metagenomic data. Methods We characterized and reported the seasonal changes of prevalent viruses and bacteria impacting the health of US Air Force Academy (USAFA) cadets, correlating sequencing pathogen identification with clinical reporting to understand the strains of various pathogens that frequently remove cadets from their duties, thereby reducing military preparedness. We employed the capture-sequencing panels VirCapSeq and BacCapSeq to test wastewater samples before, during, and after Spring Break at USAFA. An Illumina NextSeq2000 was used for sequencing and subsequent bioinformatic analyses was accomplished using a pipeline developed by GAPP. The Rapid Identification of Microbes (RIM) is a containerized pipeline that begins with trimming and quality control of raw fastq files. The web-based user interface of RIM is automatically generated and includes critical statistics to assess the trustworthiness of results, featuring links to reference

sequence and pairwise alignments, IGV reads mapping, and a Krona plot metagenomic taxonomic visualization. Results An increased number of pathogens were detected in wastewater after cadets returned from Spring Break. We identified full-length genomes of various viruses peaking at different times correlating with seasonal concerns of gastrointestinal pathogens. Conclusion: Molecular epidemiologic analysis revealed relevant mutations as the viruses evolve and persist across seasons, and phylogenetic tools such as IQtree and BEAST characterized their evolution over time. Combining clinical data with the pan-pathogen surveillance of wastewater provides leaders a powerful tool to track warfighter readiness and can be applied to public health preparedness in any location.

Oral Abstract Presentation #5: **Mutational signature analysis of whole genome sequences reveals DNA mismatch repair-deficiency in non-Cystic Fibrosis Pseudomonas aeruginosa**

Presenter: **Kalen Hall**, Ph.D., Informuta, Inc.

Abstract Body: Background: Hypermutator Pseudomonas aeruginosa, predominately caused by deficiency in DNA mismatch repair (MMR), comprises up to 60% of isolates from Cystic Fibrosis (CF) patients. MMR-deficient isolates are highly clinically correlated with multidrug resistance (MDR) and the rapid emergence of resistance in vitro. Outside of CF, the prevalence of MMR-deficiency is reported to be <1%, and the significance of MMR-deficiency in driving MDR phenotypes remains unclear. Previously, we applied computational analysis of whole genome sequences (WGS) of lab-evolved MMR-deficient P. aeruginosa to determine the resulting trinucleotide mutational signature, which we characterized as CT in NCC and NCG contexts and TC in CTN and GTN contexts. This signature was previously validated in known MMR-deficient CF clinical isolates. Methods: Here, we performed mutational signature analysis of WGS data from 392 isolates across two retrospective collections from ventilator-associated pneumonia, non-ventilated pneumonia, urinary tract infections, and intraabdominal infections to assess the prevalence of MMR-deficiency in non-CFP. Aeruginosa infections. Isolates were quantitatively defined as MMR-deficient if the cosine similarity of the spectra to signature was above 0.78, a cutoff previously determined by comparison of known MMR-deficient and MMR-proficient CF isolate spectra to the MMR-deficient signature. Results: The prevalence of MMR-deficiency varied across clinical syndromes: 9.8% in ventilator-associated pneumonia (n = 10 out of 102), 5.5% in non-ventilated pneumonia (n = 6 out of 109), 2.9% in urinary tract infections (n = 2 out of 70), and 2.7% in intraabdominal infections (n = 3 out of 111). These observed proportions are 3 to 10-fold higher than previous reports, suggesting MMR-deficiency could be contributing to MDR phenotypes and the evolution of resistance in non-CF patients. Conclusions: MMR-deficiency is seemingly enriched in ventilator-associated pneumonia, an infection exclusive to critically ill patients. We propose that this could be due to co-selection for hypermutator phenotypes arising from repeated antimicrobial treatments in such patients. Future studies aim to assess if predicted MMR-deficient isolates are correlated with MDR emergence. Rapid detection of MMR-deficient isolates using WGS and mutational signature analysis may critically aid in stratifying appropriate therapy against P. aeruginosa infections.

Scientific Session 4: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

- **Tuesday, October 15, 2:00 PM – 2:45 PM**

Oral Abstract Presentation #1: **Avian Influenza A(H5N1) Virus Among Dairy Cattle, Texas**

Presenter: **Judith Oguzie**, Ph.D., University of Texas Medical Branch

Abstract Body:

Background: H5N1 clade 2.3.4.4b emerged in late 2021 in North America. Following this, there have been numerous outbreaks and reported mammalian spillover events involving a broad range of species, with three documented cases of human infections in the US. On the 19th of March, we investigated what, at the time, was an outbreak of unknown etiology infecting dairy cows on a Texas farm. Signs observed in sick cows included decreased appetite, lethargy, increased respiratory secretions, high temperatures, abnormal bowel movements and decreased milk production. Farm workers reported influenza-like symptoms with no case of conjunctivitis, severe disease or hospitalizations. Methods: We received 40 swab specimens from 30 cows (24 sick and 6 healthy). On receiving samples, we performed molecular screenings for six viral families: adenoviruses, coronaviruses, enteroviruses, influenza viruses, paramyxoviruses, and pneumoviruses. We then used cell culture, sanger sequencing, and metagenomics sequencing on a few cohorts based on molecular screening results. Results: 29% of sick cows were H5 avian influenza A virus positive by PCR. None of the healthy cows were influenza A positive. Sanger sequencing of PCR amplicons from the HA cleavage site indicated these were HPAI. We successfully grew the virus using the three cell lines: Vero, MDBK, MDCK, and egg inoculation. Next generation sequencing confirmed HPAI in a cultured sample with successful genome assembly for the eight segments of the virus designated A/cattle/Texas/56283/2024. NGS data analysis indicated that this virus belongs to the clade 2.3.4.4b and the 4:4 reassortant genotype B3.13. In addition, we identified several mutations, including the rare M2-V27A mutation associated with adamantane resistance. Furthermore, we identified several virulence-associated mutations, including PB2 mutations (V495I and M676A) and NS1 (A223E). We also identified the PB2-M631L, which is now reported in a human HPAI sequence from Michigan and is responsible for increased viral replication in mammalian cells. In addition, we identify HA mutations N110S and V226A associated with host specificity shift and L131Q and T156A mutations responsible for antigenic drift/escape. Conclusions: The ongoing multistate multispecies outbreak of HPAI in the US underscores the importance of a one-health approach to outbreak investigations and thorough surveillance at the animal-human and environment interface to understand the complexities involved in understanding emerging diseases.

Oral Abstract Presentation #2: **Blurry *Bacillus* Boundaries: Using pangenomics to improve species delineation within *Bacillus cereus sensu lato***

Presenter: **Vishnu Raghuram**, Ph.D., Umea University

Abstract Body:

Background: Bacillus cereus sensu lato or the "B. cereus group" is a complex of Gram-positive, spore forming bacteria. Members of the B. cereus group include, among others: B. anthracis - the causative agent of anthrax; B. cereus sensu stricto - a foodborne pathogen; and B. thuringiensis, a biocontrol agent. Species-level taxonomic classification of B. cereus group strains is essential from both a public health and industrial perspective, as different strains are subject to varying forms of treatment and/or regulation. However, species identification in the B. cereus group is currently a challenge, as there is no universally accepted taxonomy. Biological and environmental barriers can influence gene gain/loss events and selection for specific genes within a species, hypothetically leading to each species unit having a distinct gene content "signature". However, these boundaries can be blurred in highly recombinogenic species. Methods: Current taxonomic classification methods use specific marker genes (e.g., panC phylogenetic group assignment, multi-locus sequence typing) or average nucleotide identity (ANI)-based thresholds. These methods vary in the number and composition of the species they define (12 - 58 different "species"). Here, we began with 5,976 B. cereus group genomes from the BTyperDB database (www.btyper.app) and used the pangenome investigation tool Panaroo to estimate the core and accessory genomes of a nonredundant subset. We then categorised genes based on distribution in the population (lineage specific, multi-lineage, all lineages), genomic context (chromosome, plasmid, phage), and functional classes (antimicrobial resistance genes, virulence factors, COG categories). Results: In this study, we sought to examine the pangenome of the B. cereus group and use differences in gene content to identify potential interspecies boundaries. We identified genetic markers fixed in specific B. cereus group lineages as well as several genes involved in potential inter-lineage recombination events, which could aid in epidemiological studies and outbreak investigations. We also discovered distinct patterns of phylogenetic dispersion of genes associated with the specific categories mentioned in "Methods". For example, we found a greater number of phage associated genes in the "lineage specific" category. Conclusions: Overall, this work provides a deeper understanding of the genomic diversity within the B. cereus group.

Scientific Session #4 Continued: Secret Ingredient: Oral Abstract Presentations

- **Tuesday, October 15, 5:15 PM – 6:30 PM**

Oral Abstract Presentation #1: **Phenotypic and genotypic evaluation of antimicrobial resistance of Mannheimia haemolytica isolated from bovine clinical samples**

Presenter: **Selma Burciaga**, Ph.D., USDA-APHIS, NAHLN

Abstract Body:

Background: Mannheimia haemolytica is a gram-negative opportunistic pathogen often inhabiting the upper respiratory tract and nasopharynx of healthy cattle. However, M. haemolytica can invade the lower respiratory tract of immunocompromised cattle following viral infections, stress, or exposure to other risk factors, playing a major role in the development of bovine respiratory disease (BRD) and enzootic pneumonia in cattle. Antimicrobial resistance (AMR) to commonly used antibiotics for treatment of BRD-associated M. haemolytica infections has been increasing in both dairy and beef cattle. The aim of this study

was to investigate AMR phenotypes and genotypes exhibited by M. haemolytica isolates obtained from bovine veterinary clinical samples during the National Animal Health Laboratory Network (NAHLN) AMR Pilot Project (2018-2022) and evaluate the suitability of whole genome sequencing (WGS) for monitoring AMR in M. haemolytica. Methods: M. haemolytica isolates were selected by participating NAHLN laboratories from routine diagnostic case submissions and tested for antimicrobial susceptibility (AST) using the Sensititre™ broth microdilution platform and commercially available Sensititre™ microdilution plates (BOPO6F and BOPO7F). WGS was performed on a subset of M. haemolytica isolates to identify AMR genes/point mutations and analyze concordance with AST results. A previously developed single nucleotide polymorphism (SNP) genotyping system was used to classify M. haemolytica into genotypes with identified associations to diseased cattle lungs, AMR genes, and integrative and conjugative elements (ICE). Results: Of 847 isolates sequenced, 280 (33%) were resistant to at least one antibiotic (mainly of tetracycline class) with corresponding AMR genes, 530 (63%) were pan-susceptible with no AMR genes, and 37 (4%) had mismatching AMR phenotypes/genotypes. Additionally, 179 (21%) isolates were multidrug-resistant with the majority harboring AMR genes aadA, ant(2'')-Ia, blaOXA-2, mph(E), msr(E), and tet(H). The M. haemolytica in our study primarily belonged to genotype 2 (87%, 737/847), a group previously associated with BRD and the presence of AMR genes. Importantly, 272 of 280 (97%) isolates harboring ≥1 AMR gene fell within genotype 2. Conclusions: Using a SNP genotyping system allows for identification of AMR and clinically relevant M. haemolytica isolated from veterinary diagnostic cases and may provide insight on mitigation strategies to prevent or reduce the prevalence of BRD on the farm.

Oral Abstract Presentation #2: **Standardized pan-genomic workflow demonstrates genomic variation of *C. sakazakii* across different isolation sources**

Presenter: **Mairui Gao**, Ph.D., University of Maryland

Abstract Body:

Background: Foodborne illness outbreaks are associated with a variety of sources, reflecting the broad niche range for many foodborne pathogens. To understand genetic signatures of pathogens at critical points along the agricultural continuum to consumers, we propose to develop a reproducible and automated workflow to identify genes and encoded pathways associated with respective isolation sources. In our study, Cronobacter sakazakii was used as a model organism to establish the workflow. Methods: Genome assemblies for C. sakazakii isolates were obtained from NCBI (n=2,089). Based on reported metadata, a fine-tune model trained with GPT was employed to automatically assign three primary isolation source types (i.e., “food”, “environment”, or “clinical”), and nine sub-category types that were relevant to this pathogen (e.g., “powdered food”, “processing facility”, “blood”). Assemblies were processed for quality control with CheckM and annotated with Prokka. A pangenome was built for high-quality assemblies (n=748) with Panaroo, and the encoded metabolic features were identified with eggNOG-mapper2. Machine learning algorithms were applied to identify accessory gene enrichments by source types and geographic locations. Results: The fine-tune model categorized all isolation sources to the primary categories with 99% accuracy and 99% precision and to the sub-categories with a 90% accuracy and 83% precision. The numbers of predicted genes per genome were significantly different among isolation sources (Clinical > Environment > Food) (P<0.05). Within sub-categories, isolates from powdered food had significantly higher genes per genome than those from

other food sources (P<0.05). Based on gene presence/absence, isolates from the different sources clustered apart along a region-dependent gradient, indicating that source type associates with genomic variation of C. sakazakii. Moreover, a random forest model predicted that the enrichments of functions associated with specific sources were largely involved in regulatory systems, metal-binding ability, heavy metal resistance, ATPase, efflux system, outer membrane lipoprotein, response regulators, etc. Conclusions: This workflow can categorize a large variety of isolation sources of C. sakazakii and identify key accessory genes and features among source types, thus suggesting mechanisms involved in adaptation and transmission. Our bioinformatics approach has broader applications for other foodborne pathogens in the future.

Oral Abstract Presentation #3: **Prediction of Clinical Outcome of *Escherichia coli* O157:H7 Infection Using Explainable Machine Learning**

Presenter: **Julian Paganini**, Ph.D., Utrecht University

Abstract Body:

Background: Shiga toxin-producing Escherichia coli O157:H7 (STEC) is a globally dispersed zoonotic pathogen, causing a range of symptoms from mild to bloody diarrhoea and haemolytic uremic syndrome (HUS). STEC's primary virulence factor, Shiga-toxin (Stx), induces host ribosome inactivation, leading to cell apoptosis. Although the Stx subtype is a key predictor of disease severity, differences in virulence with the same Stx profile are often observed. Here, we compared the performance of multiple machine learning (ML) classifiers in predicting disease severity based on whole genome sequencing (WGS) data, and evaluated the most important genomic features contributing to these predictions. Methods: WGS data from 1030 STEC isolates in the UK, reported as bloody diarrhoea (BD) (n= 597), diarrhoea (D) (n=387) and HUS (n=44), were divided into training (n=817) and test (n=213) sets, stratifying by clinical outcome and population structure. DNA kmers of varying lengths were binary encoded and used as predictive features. Feature selection involved Chi-square tests and MUVR. A final set of 1665 features were used to build the models. Random Forest (RF) and Extreme Gradient Boosting (XGB) classifiers were trained and optimised using 10-fold cross-validation stratified by population structure. Random upsampling and SMOTE were considered for dealing with class imbalance. Evaluation on the test set revealed the final model's performance. Feature importance was assessed using SHAP values. Key features were clustered based on their co-occurrence frequency. BLAST alignments against an annotated pangenome and a Stx-phage database were performed to determine the origin of each feature. Results: The overall performance of the RF and XGB classifiers was comparable, with accuracies of 0.73 and 0.72, respectively. However, the RF model had a recall of 0 for the HUS class, while XGB achieved a recall of 0.44 (Fig 1A). Upsampling did not improve classification performance in either case. SHAP values extracted from the XGB classifier revealed that feature clusters exhibiting the greatest predictive power mainly originated from genomic elements encoded by prophages and Stx-carrying phages (Fig 1B). Conclusions: XGB outperformed RF when classifying the most severe disease outcome (HUS). Feature importance analysis with SHAP revealed known and potential new predictors of disease severity.

Oral Abstract Presentation #4:	***Salmonella* Subtyping Using Hybridization (SalSUH), a highly sensitive, targeted sequencing method for identifying multiple serovars of Salmonella enterica from a metagenomic sample**
Presenter:	**Amanda Windsor**, Ph.D., U.S. Food and Drug Administration
Abstract Body:	Background: Salmonellosis linked to contaminated foods linked to over a million illnesses annually. Untargeted metagenomic sequencing for Salmonella detection is best suited for samples with a single serotype present at high relative abundance. A metagenomic method for identifying multiple Salmonella serotypes in a sample is critical to advancing food safety. Here we introduce a novel targeted capture panel, Salmonella Subtyping Using Hybridization (SalSUH), for identification of multiple Salmonella serotypes in metagenomics samples. Methods: The SalSUH panel covers 34% of the Salmonella pangenome with 10000 Salmonella-specific baits and 3367 baits diagnostic for 30 outbreak-related serotypes. Experimental cocktails with roughly equival densities of one to five Salmonella enterica serotypes and/or Salmonella bongori were spiked into a Salmonella-negative soil microbial background at high (3.6–12%), low (0.2–4.2%), or ultralow (0.02–0.43%) abundance. DNA libraries were prepared in triplicate and sequenced with and without hybridization to the SalSUH panel. All sequences were analyzed with bettercallsal v0.7.0. Results Hybridization to the SalSUH panel improved detection of S. enterica in all samples at all spike-in levels, but the most striking improvements were observed in ultralow spike-in samples. Salmonella bongori was not detected because the NCBI Pathogen Detection database, which underpins bettercallsal, does not currently have S. bongori isolate assemblies. In single serotype cocktails, at all levels, the correct serotype was detected with and without hybridization. However, hybridization resulted in an up to 422-fold increase in the percent reads aligned to the Salmonella genome. At high and low levels, up to five and four serovars, respectively were identified without hybridization; percent reads aligned ranged from 0.1–1.7%. In unhybridized, ultralow, multi-serotype samples only one serotype was identified per sample and only 0.1% reads aligned. In contrast, the same libraries hybridized to the SalSUH panel resulted in identification of the expected number and type of each S. enterica in the single cocktail. In a 5-serotype cocktail, 8.6%–14.9% reads were aligned to each serotype. Conclusions: The SalSUH targeted capture panel paired with bettercallsal allows unprecedented accuracy and sensitivity to identify multiple S. enterica serotypes in metagenomic samples. This is achievable because hybridization increases the percentage of S. enterica reads compared to those from background biota.

Scientific Session #5: Pipe Dreams: Analytical Methods, Bioinformatic Tools, and Pipelines

- **Wednesday, October 16, 9:15 AM – 10:00 AM**

Oral Abstract Presentation #1:	**It's the Yeast I Could Do: Saccharomyces Cerevisiae as a Model for Accurate Detection of RNA modifications with Direct RNA Sequencing**
Presenter:	**Kaylee Watson**, University of Maryland, Baltimore

Abstract Body:

Background: The growing field of epitranscriptomics requires novel analysis techniques to handle the complexity of data containing information about RNA modifications. Advances in direct RNA sequencing with Oxford Nanopore Technologies (ONT) allow for the detection of modifications in a native context. Select ONT basecalling models now contain expected ionic current levels for both canonical and modified bases, making RNA modification detection possible at the nucleotide-level. However, currently the only model available for RNA is specific for the m6A modification. Considering over 100 types of RNA modifications exist, it is important to be able to detect the presence of modifications other than m6A.MethodsHere we use ONT direct RNA sequencing data from two strains of Saccharomyces cerevisiae with a known differential m5C on the 25S rRNA to develop an analysis strategy for modification detection. Using the Uncalled4 software to extract position-based ionic current information, we developed a normalization approach to account for differences in sequencing depth across an RNA molecule. The normalized ionic current mean and KS statistic per position were combined and compared between the two S. cerevisiae strains to accurately predict the differentially modified position. To test this approach with varying depths and modification percentages, we generated subsets of the S. cerevisiae data using different combinations of reads from the two strains. Results: We found that a depth of 50 reads is sufficient to detect positions that are over 60% differentially modified, but with 500 or more reads, modifications at much lower percentages are easily detected above noise. As expected, no differential modification was detected between two subsets generated with reads from a single sample, which should have 0% differential modification. Conclusions: Our S. cerevisiae data has enabled us to develop a more accurate transcriptome-wide RNA modification detection strategy that can be applied to a variety of organisms.

Oral Abstract Presentation #2: **Expanding Bactopia as a framework for bacterial, fungal, and metagenomic analysis**

Presenter: **Robert Petit**, Ph.D., Wyoming Public Health Laboratory

Abstract Body:

Bactopia is a Nextflow pipeline for the complete analysis of bacterial genomes. Bactopia allows users on all types of systems such as laptops, HPC, and cloud platforms to quickly analyze bacterial genomes using hundreds of bioinformatic tools with Conda, Docker, and Singularity. Bactopia maintains extensive version control, data auditing, and testing to ensure reliability for its users. Bactopia can be rapidly scaled from a single sample to tens of thousands of samples, making it ideal for large-scale ongoing genome surveillance projects. After more than 5 years, Bactopia is one of the few analysis pipelines that has stood the test of time when it comes to consistent updates backed by user feedback. Here at the Wyoming Public Health Laboratory, we are regularly testing the adaptability of Bactopia to meet our changing needs. While Bactopia is an extensive pipeline for the analysis of bacterial genomes, at its core it is just a bunch of modules, or “building blocks'', for genomic analysis. We are now able to utilize these “building blocks'' as a framework for genomic analyses. To demonstrate this, we have developed a few new workflows with Bactopia and applied them to fungal and metagenomic sequences. Utilizing existing modules, we developed “mycotopia” for haploid fungal analysis and “clean-yer-reads”' for rapidly cleaning Illumina and ONT reads. The outputs of these workflows are compatible with the more than 60 Bactopia Tools that are available. Similarly, we used existing host-removal modules in Bactopia to develop “Teton”, which we can apply to all our sequencing to ensure all human reads are removed

prior to any analysis or public repository submission. While we still use Bactopia for our bacterial surveillance at WPHL, we have reshaped it to also act as a framework for fungal and metagenomic analysis. This allows us, as well as many from academia and public health agencies around the globe, to continue to use Bactopia for other sequencing. More importantly, these workflows require no additional training for those already familiar with Bactopia.

Oral Abstract Presentation #3: **Genome-resolved metagenomics of human gut microbiota using new long-read sequencing methods**

Presenter: **Daniel Portik**, Ph.D., PacBio

Abstract Body:

Background: The human gut microbiome contains a diversity of microbes that potentially impact health and disease. Despite considerable efforts to catalog the microbial diversity present in the human gut, large numbers of species and strains remain uncultured and undetected. Long-read sequencing is a powerful approach for resolving genomes from complex microbiomes, including the human gut. Methods: We performed a deep-sequencing experiment using PacBio HiFi reads to obtain metagenome-assembled genomes (MAGs) from a pooled human gut microbiome. We performed long-read metagenome assembly using two newer methods (hifiasm-meta, metMDBG), used improved bioinformatic and proximity ligation (Hi-C) binning strategies to cluster contigs and identify MAGs, and developed a novel framework to compare and consolidate MAGs. Results: In total, from 255 Gbp of total HiFi data we produced 595 total MAGs (including 175 high-quality MAGs) using hifiasm-meta, and 547 total MAGs (including 277 high-quality MAGs) with metaMDBG. We found proximity ligation binning yielded more MAGs than bioinformatic binning, but our novel comparison framework resulted in higher MAG yields than either binning strategy individually. Approximately 85% of the MAGs were assigned to known species, but we recovered >35 high-quality MAGs that represent uncultured diversity. Finally, we detected more total viral sequences in the metaMDBG assembly versus the hifiasm-meta assembly (~6,700 vs. ~4,500). Conclusion: We find the use of HiFi sequencing, improved metagenome assembly methods, and complementary binning strategies is highly effective for rapidly cataloging microbial genomes in complex microbiomes.

Scientific Session #5 Continued: Pipe Dreams: Oral Abstract Presentations

- **Wednesday, October 16, 10:15 AM – 11:15 AM**

Oral Abstract Presentation #1: **TB Portals Genomic Sequencing Quality Control Pipeline**

Presenter: **Madeline Galac**, Ph.D., Bioinformatics and Computational Biosciences Branch - NIAID

Abstract Body:

TB Portals is a multi-national scientific and applied medicine collaborative program spearheaded by the NIAID at the NIH focused on multidomain data from tuberculosis patients. TB Portals collects patient-centric multi-domain data,

including socioeconomic, clinical, radiological, and pathogen genomic data. Mycobacterium tuberculosis whole genome sequencing data facilitates the investigation of the epidemiological and evolutionary history of TB and enables rapid detection of drug-resistant infections and personalized therapy for patients. The TB Portals Genomics project has built a robust sequencing quality control pipeline to overcome the inevitable variability in sample quality and sequencing coverage. The majority of samples used for sequencing are collected from patient sputum, so the first part of our pipeline is a two-step process to remove any contaminating human sequences from the raw Illumina reads. The first removal step uses Kraken2 with a human genome database. Reads that are not designated as human are then mapped to a human genome reference using BWA for a secondary check with the unmapped reads continuing in our analysis. The reads are trimmed for quality and sequencing adaptors using trimmomatic. Total read counts before and after trimming are determined to provide a metric of the sequencing quality. Kraken2 is again used on the trimmed reads, now employing a database that includes all NCBI RefSeq bacterial genomes. Samples that have more than 5% of their reads from a non-TB organism are considered contaminated and are discarded. The final step involves mapping the reads to the M. tuberculosis reference H37Rv to check for genome completeness and sufficient read depth for single nucleotide polymorphism (SNP) calling confidence. For a sample to pass this step, greater than 90% of its reads need to map to the reference and cover more than 95% of the reference genome with 10X reads coverage. Sample information from our QC pipeline is shared with the corresponding microbiological labs, so they can adjust sample collection and preparation steps, if necessary. We found that such feedback leads to marked improvements in the quality of samples and genomes. Our pipeline provides valuable information to hospitals and researchers about drug-resistant and drug-sensitive tuberculosis, as well as other bacterial pathogens in clinical samples. With the help of the described pipeline, the TB Portals program has analyzed over 6,500 M. tuberculosis genomes, sequenced from over 6,000 patients in 19 countries.

Oral Abstract Presentation #2: **Generalizing Viral Genomics: Moving past one Pipeline Per Virus**

Presenter: **Daniel Park**, Ph.D., Broad Institute

Abstract Body: While the COVID-19 pandemic may have expanded viral genomic surveillance in public health labs, most of these surveillance efforts remain stubbornly single-target. Genomic sequencing of SARS-CoV-2, MPXV, RSV, influenza / HPAI, and many other viruses are still frequently performed with species-specific amplicon primers and analyzed with species-specific bioinformatic pipelines. Although mature pathogen-agnostic and multi-pathogen genomic approaches do exist, their widespread adoption in public health is held back by 1) the lack of demonstrated use cases, 2) the suboptimal (slower, often more expensive) laboratory workflows, and 3) reduced familiarity with general-purpose informatic tools. As a result, molecular and genomic surveillance remains limited to a subset of causative agents of disease. Furthermore, labs can be slow to respond to new viral threats as they take time to research, adopt, and validate new laboratory and analytic methods for each new species of virus. Our consortium recently began an OAMD-sponsored project to establish pan-pathogen metagenomic respiratory disease surveillance in a clinical setting. This project employs traditional molecular diagnostics, multi-target molecular panels, and metagenomic sequencing to understand the dynamics of circulating human respiratory viruses each winter. Over 800 clinical specimens

were sequenced in the 2023-2024 winter season, with genomes assembled spanning over 70 viral species. Here, we describe the generalized bioinformatic pipeline used to automatically identify and assemble viral genomes, from viral taxa including high diversity species, low diversity species, -ssRNA, +ssRNA, DNA, and multi-segment genomes. It is able to select closest NCBI references to assembled genomes, handle coinfections of unrelated viruses, and perform typing and subtyping for relevant species. We compare performance and outputs to some commonly used single-taxa pipelines and conclude that pathogen-agnostic methodologies--both laboratory and computational--are at least as effective at identifying and characterizing viruses of interest, and that pan-pathogen methods should be adopted more aggressively by labs engaged in genomic surveillance to speed up response times to emerging threats and enable a more representative view of circulating disease.

Oral Abstract Presentation #3: **Soft-core SNP Filtering: A New Best Practice for Bacterial Phylogenomics**

Presenter: **Mona Taouk**, The University of Melbourne

Abstract Body: Background: Phylogenomic analyses are crucial components of microbial genomics studies and can illuminate evolutionary patterns, disease transmission and outbreak dynamics. Phylogenies for bacterial populations are most commonly inferred from single nucleotide polymorphisms (SNP) alignments. These SNP alignments are typically then reduced to a ‘strict core’ that only retains sites present in 100% of the samples. However, as sample number or diversity increases, the number of sites in a strict core decreases, resulting in lost data. Hence, this approach is not optimal for large contemporary datasets that may comprise thousands of isolates. In this study, we evaluate the effectiveness instead of using a ‘soft core’ which tolerates some amount of missing data and therefore preserves more information for phylogenomics. Methods: Using publicly available large datasets of Neisseria gonorrhoeae and Salmonella enterica, we generated genome SNP alignments and assessed the number of informative sites at varying core thresholds encompassing both strict- and soft-core approaches. In a secondary analysis, we used temporal signal to quantify the accuracy of phylogenies built from strict- and soft-core alignments. To do this, we used four publicly available datasets (154 S. pneumoniae genomes, 192 N. gonorrhoeae genomes, 123 S. aureus genomes and 198 S. enterica genomes) with strong temporal signals and built phylogenies using a range of core thresholds. Results: We demonstrate that strict cores greatly reduce the number of informative sites compared to soft cores. For instance, in S. enterica, the 100% strict-core alignment of 10,000 genomes resulted in 6,365 informative sites, whereas a 95% soft core gave 64,600 informative sites. Similar trends were observed in N. gonorrhoeae, highlighting the increased resolution possible by including sites that are absent in some genomes. Our secondary analysis revealed that soft cores are typically superior to strict cores for temporal analyses. For example, the N. gonorrhoeae 95% soft-core phylogeny had a root-to-tip regression R^2 of 0.50, while the 100% strict-core R^2 was only 0.21. Conclusions: This study highlights the need to consider soft-core strategies as potential best-practice for large and diverse microbial datasets. To streamline processing, we developed Core-SNP-filter (https://github.com/rrwick/Core-SNP-filter), a publicly available resource-efficient tool implemented in Rust, that is able to provide variable and core sites based on user-supplied thresholds.