Control Number:	2022-A-12-NGS
Topic 1:	Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines
Topic 2:	
Publishing Title:	Online quality control of SARSCov-2 data with RonaQC
Author Block:	**N. Alikhan**, A. J. Page; Quadram Institute Bioscience, Norwich, UNITED KINGDOM.
Abstract Body:	Lineage assignment based on high throughput sequencing has become a key method of identifying SARS-CoV-2 VOCs/VOIs. Detecting the determining mutations is dependent on adequate upstream quality control. Spurious variant calls may be caused by cross contamination between samples, barcode cross-talk, primer contamination, primer dropouts, and carry-over from previous sequencing runs. During our evaluation of ARTIC SARS-CoV-2 Multiplex PCR detection levels and our sequencing of 15,000 SARS-CoV-2 genomes on Illumina platforms, we have identified that Illumina sequence data in particular can suffer from these artefacts. Some issues can be resolved with simple filtering, while others would render the sequencing as unreliable. These issues would not be detected by the standard QC described in the literature. Yet significant numbers of genomes available on GISAID were sequenced on Illumina platforms, and many organisations have approached us to assist them developing their SARS-CoV-2 sequencing programs. From these interactions it is clear that there is no formalised and comprehensive quality control process for SARS-CoV-2 sequencing based on Illumina data. Here we present the web resource, RonaQC, which accepts mapped SARS-CoV-2 reads (BAM format), generated from the SARS-CoV-2 bioinformatic pipelines like ARTIC, and any control samples from the respective sequencing run as input. It assesses the levels of cross contamination and primer contamination in the samples, and determines if the samples are reliable for detecting SARS-CoV-2, phylogenetic analysis, and/or submission to public databases. Reported metrics will include standard QC metrics (e.g. % of genome recovered) and metrics such as; variants should not be detected in negative control samples. Samples should have relatively significantly higher coverage than their respective negative controls. Primer contamination can be detected by the detecting fragmented, or partially mapped reads, in localised high coverage genome regions. The output to the users includes a simple traffic light system assessment for non-expert audiences with links to additional detailed graphs and metrics to peruse. The web resource output is designed to be easy to understand and largely graphical like other QC tools such as FASTQC. The web resource was implemented through HTML+javascript using libraries such as react.js (user interface) and d3.js (visualisation). The assessment of read mapping output (BAM) required the use of samtools that are run within the user's web browser with webassembly. This means, firstly, that the application runs entirely through a web browser and requires no software installation on the user's device. And secondly, the user's sequencing data remains entirely on their local computer, which is a consideration for those in public health with restricted data sharing requirements.

---PAGEBREAK---

Control Number: 2022-A-13-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Advantages of Hybrid Versus Short-Read Only Assembly for Resolving Prokaryotic Genomes

Author Block: **E. Cella**, M. Jubair, T. Azarian;
Burnett School of Biomedical Sciences, University of Central Florida, ORLANDO, FL.

Abstract Body: The advent of 3rd Generation ultra-long read sequencing platforms such as those developed by Oxford Nanopore Technologies (ONT) as well as commensurate improvements to computational methods for hybrid genome assembly have made routine closure of prokaryotic genomes feasible. However, the advantage of generating complete (i.e., circularized) hybrid assemblies versus draft assemblies from short-read sequencing data alone has not fully been explored. Here, we systematically compared draft bacterial genomes of 90 *Staphylococcus aureus* (SA) strains assembled using short-read (Illumina) only sequencing data to assemblies generated with hybrid (Illumina + ONT) data. Assemblies were performed using Unicycler v0.4.8 and annotated with Prokka v.1.14.6 and NCBI Prokaryotic Genomes Annotation Pipeline (PGAP). Pangenome analysis was performed using Roary v.3.12 and phylogenies were inferred using IQtree v2.0.3. Assembly, annotation, pangenome, and phylogeny statistics were compared between short-read only and hybrid assemblies. Hybrid genome assembly fully resolved and circularized 73/90 (81%) SA genomes, which included complex genomes possessing multiple plasmids. The remaining 17 genomes were assembled into linear contigs (n=6) or had ≤11 unatigs (n=11). The mean assembly length for the hybrid assemblies was ~2.8 Mbps, which was an average of 46 Kbps greater than short-read only assemblies. PGAP annotation of hybrid assemblies assigned on average 2,704.6 coding sequences (CDS). Core genome SNP alignments obtained from hybrid assemblies annotated with PGAP had comparable parsimony informative sites as compared to short-read only assemblies, 79.06% vs 79.05% respectively. Phylogenies inferred from the core genome alignment of hybrid assemblies were more resolved than those inferred from short-read draft genomes based the scaled length of terminal branches as well as assessment using the Kendall-Colijn metric. We find that hybrid assembly is a highly effective approach to resolving prokaryotic genomes. Yet, short-read only assemblies yielded comparable results in terms of annotation, pangenome analysis, phylogenetic signal, and tree topology. The decision to more completely resolve bacterial genomes using hybrid depends on the goal of the analysis. Hybrid approaches often fully resolve mobile genetic elements such as plasmids and provides enhanced resolution that is important for assessing genome organization (i.e., synteny), content (e.g., pangenome), recombination history, and micro-evolution. As 3rd generation sequencing technology continues to improve, long-read only assembly may achieve accuracy negating the need for short-read data. This would improve the cost and feasibility of routine closure of prokaryotic genomes, which may provide an enhanced understanding of the ecology and evolution of a bacterial organisms.

---PAGEBREAK---

Control Number: 2022-A-15-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Finding New and Unique Drug Targets against the Resistant Microbial Pathogens using Bioinformatics Pipeline

Author Block: **R. Uddin**;
University of Karachi, Karachi, PAKISTAN.

Abstract Body: The rapid increase in the resistance of microbial pathogens pose serious threats to human life. The W.H.O. has placed such resistant bacteria among top health challenges faced by humans. The pathogen's normal metabolic mechanism disturbs by the drug or antibiotic which is administered to cure the disease. However, the pathogen has the capability to bypass and achieve new and unique pathways of survival in that acute unfavorable condition and hence resulted in resistance. The research and development to propose new therapies are imperative need of time. Discovery of novel drug candidates is not an easy task in the early stages of drug discovery, at the same time expensive and laborious. On top of that, the time requires to bring a drug in an open market is quite long (~14 years). Identification of unique drug targets, on the other hands, is a routine method these days, particularly for the resistant pathogens. In this method, we try to prioritize the novel and unique drug targets among the druggable genomes of the infectious pathogens. In this way, we may bring the novel and unique drug targets that are essential to bacterial survival and against which the new drugs can be discovered. In this talk, we will share our work in this area of research and will show successful examples of using Computational Biology methods to prioritize novel and unique drug targets against resistant pathogens e.g. M. tuberculosis, Methicillin-Resistant S. aureus among others.

---PAGEBREAK---

Control Number: 2022-A-18-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: A general flexible framework to evaluate performance and provide guidelines for bioinformatics methods for taxonomic classification and resistance gene detection based on noisy long metagenomic reads

Author Block: **A. Van Uffelen**, A. Posadas, N. H. Roosens, S. C. De Keersmaecker, K. Vanneste;
Sciensano, Brussels, BELGIUM.

Abstract Body: With shotgun metagenomics, DNA directly extracted from all cells in a community is sequenced. It allows assumption-free and unbiased microbial detection based on the ability to detect any and all genomic information from various biological organisms within a sample. Especially the use of 3rd-generation sequencing technologies, through the production of long reads, enables the detection and scaffolding of microbial genes to their host chromosomes and taxonomic classification with an unprecedented sensitivity. However, several bottlenecks still need to be overcome: the high error rate of long reads, unstable protocols and tools, lack of guidelines and best practices, and unanswered questions about performance in a quality setting. We have built a general framework to evaluate the performance of taxonomic classification of the microbial composition, detection of antibiotic resistance genes, and their genetic context (*i.e.*, chromosomal or plasmidic) by reporting several relevant performance metrics. The framework is flexible and allows easily switching between methods, tools, parameters, and databases to evaluate which methods work best for a certain use case. We are currently benchmarking several long-read-based taxonomic classification methods to provide guidelines and advice for

employing them within an applied public health/clinical setting to build a resource that will also benefit other laboratories and/or application domains.

---PAGEBREAK---

Control Number: 2022-A-21-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Genome and transcriptome analysis of the microalgae *Desmodesmus abundans* used for flue gas mitigation and byproduct obtaining after 13 years under high-CO_2

Author Block: **S. Mora Godínez**, A. Pacheco;
Tecnológico de Monterrey, Monterrey, MEXICO.

Abstract Body: Major contributors to CO_2 emissions are the fossil fuel combustion and cement industry. The high CO_2 fixation rates and biomass productivity of microalgae makes them an alternative for biological mitigation of flue gases and high value byproducts obtaining. Also, other gas components such as NO_x and SO_x can be used as nutrients. The objective of the study was to analyze the transcriptome, genetic variations and biomass composition of *Desmodesmus abundans* acclimated for 13 years to high CO_2 (HCA, 50% v/v) and low CO_2 (LCA, air, 0.04%), and grown under a model cement flue gas (MFG, 25% CO_2, 700 ppm NO, 100 ppm SO_2) in a N- and S- depleted medium (BG11-N-S) using 1 L column photobioreactors. Strain HCA under MFG presented a higher rate than LCA; however, both strains presented similar biomass productivities at the end of the run (0.30-0.34 g d.w. L^{-1} d^{-1}). Transcriptome analysis showed 16 435 up- and 4 219 down-regulated contigs in HCA. These were related to synthesis of nucleotides and amino acids, and central carbon metabolisms (C3 and C4 cycles, glycolysis, TCA cycle and gluconeogenesis). Contigs assigned to cellular component GO terms (chloroplast, photosystem I and II, and cell wall) were differentially expressed, and mostly up-regulated in HCA. Similar, nitrogen transporters were up-regulated, in agreement with this HCA presented a higher consumption of nitrogen. The expression of genes involved in secondary metabolite pathways such as anabolism and catabolism of starch and triacylglycerol (TAG) were also higher in HCA. Both strains presented a similar biomass composition, with 21% w/w protein, 3-4% w/w crude fat (neutral lipids), 10-11% w/w crude fiber, and 59% w/w carbohydrates. Moreover, total lipid content (polar and non-polar) and starch were higher in LCA, and pigments in HCA. Lipidome analysis resulted in an increase in glycerophospholipids in HCA, while TAG increased in LCA. Controls using complete medium (BG-11) and high CO_2 showed a higher cell concentration and protein than MFG (BG-11-N-S), but a lower starch content probably as consequence of limited-N in MFG. In accordance, the lipidome of microalgae under MFG versus the control presented an increase in 54 TAG which are related to stress conditions, and a decrease in 33 GP that are part of membrane structure. Both strains demonstrated capacity to tolerate and use flue gas as nutrient source. However, after 13 years in high CO_2, differences in transcriptome were evident as well in byproducts obtaining (starch, pigments and lipid profile). Currently, studies continue to study strain genetic variations and to optimize growth under flue gas using continuous photobioreactors and N-supplementation to optimize CO_2 fixation and productivities.

---PAGEBREAK---

Control Number: 2022-A-27-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Supporting the bioinformatics community by supporting users of Bactopia

Author Block: **R. A. Petit III**[1], T. Fearing[1], C. Rowley[1], J. Mildenberger[1], T. D. Read[2];
[1]Wyoming Public Health Laboratory, Cheyenne, WY, [2]Emory University, Atlanta, GA.

Abstract Body: The ongoing maintenance of open source bioinformatic tools is a challenging process. All too often maintenance of these tools is time-consuming and completely voluntary with few incentives to continue support. Few funding mechanisms exist to support open source bioinformatic tools, but these are targeted at larger projects. There are currently no funding mechanisms to support many of the essential tools used in bacterial genomics, leaving it up to the community to voluntarily help support these tools. With this in mind, we designed Bactopia to help support and contribute back to the bioinformatics community.

Bactopia is an open source bioinformatic pipeline for the complete analysis of bacterial genomes. It is written in Nextflow DSL2 following nf-core practices and easily accessed through Bioconda, Docker, and/or Singularity. The main pipeline takes each sample through a standard set of analyses to produce a broad overview of samples. Using these outputs, Bactopia Tools allow users to dive deeper into their samples. In addition to bioinformatic outputs, Bactopia also produces numerous audit logs including version information.

Bactopia allows users to rapidly take advantage of more than 130 open-source bioinformatic tools. Bactopia includes mechanisms to contribute back to these tools. To include a tool in Bactopia, it must be open-source and available on Bioconda. There are many benefits to using open-source tools, but an important one is bug fixes can be submitted back to the tool. By only using tools available from Bioconda, it promotes a valuable asset to the field of bioinformatics which includes many downstream efforts such as containerization. Additionally, all Bactopia Tools must be available from nf-core/modules which allows these tools to be seamlessly added to any Nextflow pipeline.

By implementing these mechanisms, we have created an efficient feedback loop with our users. Overtime, users help identify and prioritize which tools are causing issues or those that should be included in Bactopia. We then determine if a fix is possible and submit a pull request, or if new tools need to be added to Bioconda or nf-core/modules, and if so, we add them. By directly supporting our users, we have been able to make more than 125 contributions to the wider bioinformatic community. Ideally in the future, there will be funding available to directly support on-going maintenance of bioinformatic tools. Until then, it is essential that pipeline developers develop similar relationships with their users to help support the tools they are using.

---PAGEBREAK---

Control Number: 2022-A-29-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Assessing the spatial and temporal variability of bacterial communities in two Bardenpho wastewater treatment systems via Illumina MiSeq sequencing

Author Block: **s. sherchan**;
Morgan State University, Baltimore, MD.

Abstract Body: Next generation sequencing provides new insights into the diversity and ecophysiology of bacteria communities throughout wastewater treatment plants (WWTP), as well as the fate of pathogens in wastewater treatment system. In the present study, we investigated the bacterial communities and human-associated Bacteroidales (HF183) marker in two WWTPs in North America that utilize Bardenpho treatment processes. Although, most pathogens were eliminated during wastewater treatment, some pathogenic bacteria were still observed in final effluents. The HF183 genetic marker demonstrated significant reductions between influent and post-Bardenpho treated samples in each WWTP, which coincided with changes in bacteria relative abundances and community compositions. Consistent with previous studies, the major phyla in wastewater samples were predominantly comprised by Proteobacteria (with Gammaproteobacteria and Alphaproteobacteria among the top two classes), Actinobacteria, Bacteroidetes, and Firmicutes. Dominant genera were often members of Proteobacteria and Firmicutes, including several pathogens of public health concern, such as Pseudomonas, Serratia, Streptococcus, Mycobacterium and Arcobacter. Pearson correlations were calculated to observe the seasonal variation of relative abundances of gene sequences at different levels based on the monthly average temperature. These findings profile how changes in bacterial communities can function as a robust method for monitoring wastewater treatment quality and performance for public and environmental health purposes.

---PAGEBREAK---

Control Number: 2022-A-43-NGS
Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems
Topic 2:

Publishing Title: Performance of bioinformatics tools for the detection of insecticidal protein-encoding genes in *Bacillus cereus sensu lato* biovar Thuringiensis genomes

Author Block: **T. Chung**[1], A. Salazar[1], G. Harm[1], S. Johler[2], L. M. Carroll[3], J. Kovac[1];
[1]Department of Food Science, The Pennsylvania State University, University Park, PA, [2]Institute for Food Safety and Hygiene, University of Zurich, Zurich, SWITZERLAND, [3]Structural and Computational Biology Unit, EMBL, Heidelberg, GERMANY.

Abstract Body: *Bacillus cereus sensu stricto* (*s.s.*) biovar Thuringiensis (Bt) strains belong to the same genomospecies as foodborne pathogen *B. cereus s.s.* and are from *Bacillus cereus sensu lato* (*s.l.*) group. Bt strains are identified based on the presence of *cry*, *cyt*, and *vip* genes, which encode insecticidal crystal proteins (i.e., Bt toxins). Multiple bioinformatics tools have been developed for the detection of crystal protein-encoding genes based on whole genome sequencing (WGS) data. However, the performance of these tools is yet to be evaluated using phenotypic data. The goal of this study was therefore to (i) phenotypically screen a collection of *B. cereus s.l.* strains for production of crystal proteins and (ii) assess the performance of different bioinformatics tools for the detection of crystal protein-encoding genes.
A total of 78 *B. cereus s.l.* strains isolated from foods, environmental sources, and commercial biopesticide products underwent whole-genome sequencing using illumina NextSeq. Sequencing reads were trimmed using Trimmomatic (v0.3.6), and *de novo* assembled using Unicycler (v0.5.0). Crystal protein-encoding genes were detected

using BtToxin_Digger (v1.0.10), BTyper3 (v3.2.0), IDOPS (v0.2.2), and Cry_processor (v1.0). For phenotypic detection of crystal proteins, isolates were grown on T3 agar for 72 to 120 h at 30°C and examined for crystal protein production using wet mount phase contrast microscopy. The sensitivity and specificity for predicting crystal protein production based on detected crystal-protein encoding genes (*cry*, *cyt*, and *vip*) were determined for each tested bioinformatics tool.
Out of 78 isolates, the production of crystal proteins was confirmed for 33 isolates. Specificity and sensitivity were 0.78 and 0.97 for BtToxin_digger, 0.96 and 0.88 for BTyper3, 0.93 and 0.97 for IDOPS, and 0.98 and 0.88 for Cry_processor, respectively. Cry_processor predicted crystal protein production with highest specificity, and BtToxin_Digger and IDOPS predicted crystal protein production with highest sensitivity, whereas BtToxin_digger had the lowest specificity. Three out of four tested bioinformatic tools performed well overall, and IDOPS performed best, as indicated by both high (> 0.90) sensitivity and specificity.

---PAGEBREAK---

Control Number: 2022-A-47-NGS

Topic 1: Microbial Chatter: Microbial ecology in health and disease

Topic 2:

Publishing Title: Microbial Community Distribution and Core Microbiome in Graded Buruli Ulcer and Chronic Tropical Wounds

Author Block: **M. Frimpong**[1], A. Kreitman[1], B. Agbavor[2], O. Dornu[2], R. Phillips[2], E. Ghedin[1];
[1]National Institutes of Health, Bethesda, MD, [2]Kumasi Centre for Collaborative Research, Kumasi, GHANA.

Abstract Body: **Background:** Buruli ulcer (BU) is a neglected tropical disease caused by *Mycobacterium ulcerans* infection that damages the skin and subcutaneous tissue. The disease is most prevalent in rural West and Central Africa and some part of Australia. Early antibiotic treatment gives good outcomes with varied timelines to complete healing ranging from days to several months even for wounds with similar characteristics. In most rural communities where Buruli ulcer is endemic, there is a wide range of other ulcers of unknown origin typically classified as chronic tropical ulcers. When treatment is delayed or disease is diagnosed late, severe morbidity, permanent disability, social stigma and loss of productivity can ensue. Analysis of microbial landscape is a major requirement towards devising evidence-based intervention. **Methods and Findings:** Towards this, 170 wound (90 BU and 80 tropical ulcers) samples were sampled for their bacterial community structure using 16S rRNA target gene sequencing and whole genome shotgun metagenomics. The 16S rRNA gene analysis showed that both gram-negative and gram-positive bacteria were abundant in the wound microbiome. The core microbiome consisted mostly of bacteria genera *Pseudomonas, Staphylococcus* and *Proteus* in decreasing order of relative abundance in both groups. Bacteria genera, *Corynebacterium* was differentially abundant in the tropical ulcers whilst *Aeromonas* and *Catonella* were shown to be differentially abundant in the Buruli ulcer lesions. On-going metagenomics analyses of samples in the 2 groups include species-level taxonomic assignments and characterization of antibiotic resistance genes to determine associations with treatment regimens and wound healing.**Conclusion:** The detection of polymicrobial communities in these tropical ulcers could guide effective wound management. A detailed resolution of the microbial communities to the species level and

antimicrobial resistant patterns will aid clinicians to tailor their treatment specifically to the microbes prevalent at the time of assessment.

---PAGEBREAK---

Control Number: 2022-A-49-NGS
Topic 1: Microbial Chatter: Microbial ecology in health and disease
Topic 2:
Publishing Title: Antimicrobial resistance trends among Escherichia coli and Klebsiella pneumoniae at Ethiopian Public Health Institute, Addis Ababa, Ethiopia: Retrospective analysis (2017-2021)
Author Block: **A. Abdeta**, G. Guma, A. Negeri, S. Fentaw, D. Beyene;
Ethiopian Public Health Institute, Addis Ababa, ETHIOPIA.
Abstract Body: **AbstractPurpose**: Periodical report of antimicrobial resistance trend data is important to help infection control efforts. Therefore, this study was designed to analyze five-year trends of antimicrobial resistance profiles among *Escherichia coli* and *Klebsiella pneumoniae*.**Materials and Methods**: A retrospective study was conducted to analyze the antimicrobial resistance trends among *Escherichia coli* and *Klebsiella pneumoniae* isolated from blood, pus, and urine specimens from 2017 to 2021 using data obtained from Ethiopian Public Health Institute. We grouped related antimicrobials according to their drug classes for trend analysis. R software was used to perform statistical analysis. Cochran Armitage trend test was employed to test the significance of antimicrobial resistance trends over time. The P-values < 0.05 were considered statistically significant.**Results**: A total of 5382 bacteriology culture data with complete information were included, from which 458 (8.5%) *E. coli* and 266 (5%) *K. pneumoniae* were obtained. *K. pneumoniae* had high resistance to extended-spectrum cephalosporins (88%) and the lowest resistance to carbapenems (14%). Moreover, *E. coli* demonstrated highest resistance to trimethoprim/sulfamethoxazole (75%), with the lowest resistance to carbapenems (5%). *K. pneumoniae* showed a significantly increasing resistance to carbapenems (0% in 2017 to 38% in 2021, P-value<0.001) and ciprofloxacin (41% in 2017 to 90% 2021, P-value<0.001), whereas *E. coli* showed decreasing resistance to β-lactam/β-lactamase inhibitors (P-value <0.001).**Conclusion**: The increasing resistance to last-resort antimicrobials particularly among *K. pneumoniae* suggests that antimicrobials are being misused in Ethiopia, necessitating the need for improved antimicrobial stewardship programs.**Keywords**: Antimicrobial resistance, *E. coli*, *K. pneumoniae*, Ethiopia

---PAGEBREAK---

Control Number: 2022-A-51-NGS
Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems
Topic 2:
Publishing Title: Antimicrobial Genes in Thermophilic *Bacillus paralicheniformis* Associated with Mobile Elements
Author Block: **O. Elsakhawy**, M. Abouelkhair, S. Kania, R. Jones, S. Rajeev;
University of Tennessee, Knoxville, KNOXVILLE, TN.
Abstract Body: Bacteria in extreme environments adapt rapidly and produce an array of antimicrobials to inhibit other organisms as they compete for scarce resources. Thermal features represent

one such environment and are useful to study of the development, evolution and transmission of antimicrobial genes. However, horizontal gene transfer from extremeophile bacteria has been discounted due to a lack of association with mobile genetic elements. This study examined antimicrobial resistance genes and their genomic locations relative to transposons and bacteriophage in *Bacillus paralicheniformis* isolated from hot springs. Bacteria were collected in Yellowstone National Park during 2020 (Permit YELL-2020-SCI-8152). Water temperatures in collection sites averaged 65.1 °C. For propagation, samples were divided into three types of liquid media: malt yeast, ATCC Medium 1554 (mineral salt), and peptone-yeast-glucose. Samples were incubated aerobically and anaerobically at 37°C, 60°C and 70°C for 10 days at which time turbidity was noted. Broths positive for growth were subcultured to blood agar plates and bacterial colonies were isolated, DNA was extracted and stocks were frozen. 16S rRNA genes were amplified by PCR and sequenced using the Sanger technique. Three unique isolates were identified for further study. A strategy was developed to identify difficult to place elements including transposons, bacteriophage and plasmids. We used a short-read-first hybrid assembly method (short-read assembly followed by long-read bridging and polishing). SPAdes v3.14.0 was used to assemble the Illumina Hiseq short reads into an assembly graph using a variety of k-mer sizes, evaluating the graph at each step to select the graph with the lowest contig count and dead-end count. The SPAdes contigs were then polished using Racon v1.5.0 and miniasm v 0.3 with MinION nanopore long uncorrected reads. Genes associated with resistance to bacitracin (*bcrA-1, bcrA-2, bcrA_3, bcrA_4, bcrB*), beta lactamase and its accessory proteins (*BlaR1, BlaI, penP*), virginiamycin (*vgb)*, oleandomycin (*oleD_1, oleD_2*) , bicyclomycin (*bcr_1, bcr_2*), linearmycin (*lnrL_1, lnrL_2, lnrL_4, lnrL_5, lnrN, lnrL_M*), chloramphenicol (*CAT*) and rifamycin (*rpnC_1, rpnC_2*) were identified. Multiple copies of Transposons Tn3, IS1595, IS1182, and IS200/IS605 were located in several positions. These transposons, which have been associated with horizontal gene transfer, were found in close proximity to clusters of resistance genes. These findings, as well as the identification of bacteriophage and plasmids suggest the potential for exchange of resistance genes between organisms in this extreme environment.

---PAGEBREAK---

Control Number: 2022-A-62-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: A strategy for screening broad-spectrum antiviral drugs against respiratory viruses

Author Block: **O. Bajinka**;
University of The Gambia, Banjul, GAMBIA.

Abstract Body: The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), leading to COVID-19, has become a current threat to human. Huge new shocks from SARS-CoV-2 are spreading worldwide with various new variants of it. Other respiratory viruses, including RSV, Influenza, hRV, are also streaming, which constitute major public health problems. However, the pathogenesis and the mechanism of various respiratory viruses remain unclear. Therefore, we implemented transcriptomic analysis to detect common pathways and molecular biomarkers in SARS-CoV-2, RSV, influenza and hRV that help understand the association between SARS-CoV-2, RSV, influenza and hRV patients. Alternatively,

identifications of specific pathways and cell markers also help to distinguish infection induced by different viruses. Here, four RNA-seq datasets (GSE152075, GSE97742, GSE103166 and GSE93731) from Gene Expression Omnibus (GEO) are employed to detect differentially expressed genes (DEGs) for SARS-CoV-2, RSV Influenza and hRV infection. A total of 19 common DEGs among these four datasets were identified. Using a bioinformatics approach, we constructed the protein-protein interaction (PPI) and identified Hub genes. After GO and pathway analysis of common DEGs, we found that the host significantly intranasally activated the antiviral response after infection, and we also identified the expression regulatory network of these common genes. We found that FOXC1, GATA2, YY1, NFIC as well as MEF2A are the major upstream transcription factors. CMPK2 is the top-ranked hub gene, which promotes NLRPS inflammasome signaling pathway to induce acute respiratory diseases (ARDs).Through the screening of drugs, we found that estradiol, Erlotinib, Lapachone as well as Sonidegib may serve as treatment option for COVID-19 through CMPK2 and that these drugs may have a broad spectrum of antiviral effects.

---PAGEBREAK---

Control Number: 2022-A-75-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Subtyping Evaluation of *Salmonella* Enteritidis Using SNP and Core Genome MLST with Nanopore Reads

Author Block: **Z. Xian**[1], S. Li[2], D. A. Mann[1], Y. Huang[1], F. Xu[3], X. Wu[3], S. Tang[3], G. Zhang[3], A. Stevenson[4], C. Ge[3], X. Deng[1];
[1]Center for Food Safety, University of Georgia, Griffin, GA, [2]School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou, CHINA, [3]Mars Global Food Safety Center, Beijing, CHINA, [4]Mars Advanced Research Institute, Mclean, VA.

Abstract Body: Whole genome sequencing (WGS) for public health surveillance and epidemiological investigation of foodborne pathogens predominantly relies on sequencing platforms that generate short reads. Continuous improvement of long-read nanopore sequencing such as Oxford Nanopore Technologies (ONT) presents a potential for leveraging multiple advantages of the technology in public health and food industry settings, including rapid turnaround and onsite applicability in addition to superior read length. However, evaluation, standardization and implementation of the ONT approach to WGS-based, strain-level subtyping is challenging, in part due to its relatively high base-calling error rates and frequent iterations of sequencing chemistry and bioinformatic analytics. Using an established cohort of Salmonella Enteritidis isolates for subtyping evaluation, we assessed the technical readiness of ONT for single nucleotide polymorphism (SNP) analysis and core-genome multilocus sequence typing (cgMLST) of a major foodborne pathogen. By multiplexing three isolates per flow cell, we generated sufficient sequencing depths under seven hours of sequencing for robust subtyping. SNP calls by ONT and Illumina reads were highly concordant despite homopolymer errors in ONT reads (R9.4.1 chemistry). In silico correction of such errors allowed accurate allelic calling for cgMLST and allelic difference measurements to facilitate heuristic detection of outbreak isolates. Our study established a baseline for the continuously evolving nanopore technology as a viable solution to high quality subtyping of Salmonella, delivering comparable subtyping performance when used standalone or together with short-read platforms.

---PAGEBREAK---

Control Number: 2022-A-77-NGS

Topic 1: Microbial Chatter: Microbial ecology in health and disease

Topic 2:

Publishing Title: MftP is a Multi-Drug Efflux Pump with a Vital Role in Regulating Cellular Homeostasis in *Burkholderia thailandensis*

Author Block: **A. Al-Tohamy**, F. Donnarumma, A. Grove;
LSU, Baton Rouge, LA.

Abstract Body: Misuse of antibiotics has accelerated the spread of drug resistant microbes. A key component of antimicrobial resistance is efflux pumps, which extrude hazardous compounds from the microbial cell. Major facilitator transport protein (MftP) from *Burkholderia thailandensis* is a Major Facilitator Superfamily (MFS) protein with high similarity to EmrD, a proton-dependent exporter from *Escherichia coli*. MFS efflux pumps are ubiquitous, suggesting that they may be important for detoxification of intracellular metabolites in addition to exporting antibiotics. In order to identify potential substrate(s) for MftP, *ΔmftP* and wild-type strains were grown on 2xYT plates containing different antibacterial agents. The *ΔmftP* cells were more sensitive than wild-type cells to ten different compounds, suggesting that MftP is a multidrug efflux pump. For instance, we identified benzylpenicillin, a beta-lactam antibiotic, as a substrate for MftP. To address the ability of MftP to export cellular metabolites, we conducted a mass spectrometry high throughput quantitative proteomics analysis for *ΔmftP* and wild-type strains using 6-plex tandem mass tag (TMT). Bioinformatics techniques were employed to identify enriched signal pathways based on proteins that were differentially expressed. In total, 2853 proteins were detected with a high-level confidence. In *ΔmftP* compared to wild-type, 293 proteins were differentially expressed, including 155 accumulating and 138 depleting proteins. Upregulated proteins participated in processes such as chemotaxis, flagella formation, and the type III secretion system. Moreover, four secondary metabolite pathways were upregulated, including synthesis of bactobolin, hydroxyalkyl-quinoline, malleilactone, and catechol. Proteins that were downregulated showed an enrichment in the type VI secretion system, pili biogenesis, and beta-lactam resistance pathways. The latter is consistent with the increased sensitivity of *ΔmftP* cells to the beta-lactam antibiotic benzylpenicillin. We propose that the differential production of numerous proteins reflects an accumulation of cellular metabolite(s) in *ΔmftP* cells, which in turn influences differential gene expression, and that MftP is key to maintaining homeostasis of specific metabolite(s). Additionally, our findings suggest that roles of efflux pumps in maintaining normal cellular homeostasis must be considered when pursuing efflux pumps as a potential therapeutic target.

---PAGEBREAK---

Control Number: 2022-A-80-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Metagenomic Antimicrobial Susceptibility Testing from Simulated Native Patient Samples

Author Block: L. Lüftinger[1], A. Materna[1], T. Rattei[2], **S. Beisken**[1];
[1]Ares Genetics GmbH, Vienna, AUSTRIA, [2]University of Vienna, Vienna, AUSTRIA.

Abstract Body: Clinical metagenomics, the culture-free metagenomic sequencing of native patient samples, enables the molecular characterization of bacterial pathogens and genomic antimicrobial susceptibility testing (AST) through novel bioinformatics tools.
While several in-silico methods to determine AST results from bacterial whole-genome sequencing data have been shown to be accurate for many pathogens and antimicrobials [1, 2], the applicability and performance of these methods has not been evaluated for metagenomic data. The presence of human background and contaminant species as well as the challenge of assigning antimicrobial resistance (AMR) markers to individual taxa in polymicrobial samples, may significantly degrade performance.
We use the AMR reference database ARESdb [3] to investigate the performance of bioinformatics tools for the prediction of AST results from clinical metagenomic data (MG-AST). We simulate 576 clinical metagenomes from 48 Escherichia coli and 51 Klebsiella pneumoniae from a multi-center study on AMR as well as shotgun-sequenced septic urine samples. We apply the rule-based tool ResFinder [2] and in-house machine learning-based AST classifiers [4] to investigate the impact of (a) levels of human background reads, sequencing depth, and metagenome complexity as well as (b) different metagenomic assembly/binning approaches on MG-AST accuracy.
Results indicate that, given a minimum sequencing depth to allow for 10x de-novo genome assembly, contamination by urogenital flora below 5% of total reads and presence of Homo sapiens background do not significantly affect MG-AST accuracy compared to isolate sequencing data (p < 0.05). An increase in the major error of <2% was observed; for co-infection scenarios with two closely related pathogens, standard metagenomic binning cannot reconstitute genomes suitable for tools developed for isolate sequencing data. The loss of AMR markers and plasmids during binning contributes to an unacceptable increase in the very major error rate. We explore options to characterize the resistome of metagenomes and improve the sensitivity of MG-AST from metagenomic bins.
1. 10.1128/JCM.00273-20
2. 10.1093/jac/dkaa345
3. 10.1016/j.gpb.2018.11.002
4. 10.3389/fcimb.2021.610348

---PAGEBREAK---

Control Number: 2022-A-91-NGS
Topic 1: Epidemiological Cues: NGS in Clinical and Public Health Microbiology
Topic 2:
Publishing Title: GenomeTrakr Database and Network Updates 2022

Author Block: **M. Allard**, R. Timme, M. Timme, S. Cianci, E. Stevens, M. Hoffmann, G. Kastanis, T. Muruvanda, J. Payne, A. Pightling, H. Rand, J. Pettengill, Y. Luo, N. Gonzalez-Escalona, D. Melka, P. Curry, Y. Chen, S. Tallent, E. Brown;
FDA, South Hero, VT.

Abstract Body: A network of federal, state, academic, and international laboratories has been using WGS data to rapidly characterize pathogens since 2012. Sequences from this GenomeTrakr network are curated by NCBI and available through NCBI Pathogen Detection web portal.

The GenomeTrakr database has demonstrated how distributed network of desktop WGS sequencers can be used in concert with traditional epidemiology and investigation for source tracking of foodborne pathogens. This “open data” model allows greater transparency between federal/state agencies, industry partners, academia, and international collaborators. This report documents the value that is gained from analyses of a million pathogen genomes. NCBI, currently is producing daily phylogenetic results for 49 bacterial pathogens including: *Salmonella*, *Listeria*, *E. coli* and *Campylobacter* for > 67,000 clusters of ≤ 50 SNPs, plus pairwise SNP count differences for cluster members. A second NCBI product called AMRFinderPlus provides genotype calls of presence/absence for known antimicrobial resistance (AMR) genes as well as genotypes related to virulence and stress. We discuss GalaxyTrakr for global distribution of our bioinformatic pipelines, Protocols.io for sharing methods and communication and integration. The new FDA laboratory flexible funding model has distributed funds to double the existing network of domestic partners and has recently distributed funding for SARS-CoV-2 wastewater special surveillance project associated with human food workers. Results demonstrate global benefits of having an open data model. Understanding root causes of foodborne contamination and assisting our academic, public health and industry partners to develop preventative controls to make food safer globally.

---PAGEBREAK---

Control Number: 2022-A-92-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Whole-Genome Sequence-Based Analysis of the *Bacillus* Sp. And *Corynebacterium* Sp., Two Metabolizing Bacteria Isolated from Two Major Landfills in Lagos, Nigeria

Author Block: **A. K. Ogunyemi**[1], O. M. Buraimoh[1], B. C. Ogunyemi[2], S. K. Odetunde[3], T. T. Oshin[3], T. A. Samuel[1], O. O. Amund[4], M. O. Ilori[1], O. O. Amund[4];
[1]University of Lagos, LAGOS, NIGERIA, [2]Yaba College of Technology, LAGOS, NIGERIA, [3]Lagos State University of Science and Technology, LAGOS, NIGERIA, [4]Elizade University, Ondo State, NIGERIA.

Abstract Body: Next-generation DNA sequencing (NGS) has made it feasible to sequence large number of microbial genomes and advancements in computational biology have opened enormous opportunities to mine genome sequence data for novel genes and enzymes or their sources. Therefore, this study aimed to identify and perform a genomic investigation on two nitrile-metabolizing *Bacillus* sp. strain and *Corynebacterium* sp strain. isolated from two major landfills in Lagos, Nigeria. Growth assays verified the two strains be able to metabolize glutaronitrile and benzontrile as sole carbon source, and with a preference for aerobic conditions (pH 7, 37-45 °C). WGS (Whole genome sequencing) was performed for the nitrilase-producing strains WOD8 and WOIS2; would show an array of genes involved in the nitrilase production identified in several different genomic locations, guiding potential genetic manipulation studies in the future. The whole-genome sequence-based analysis would reveal that strains belonged to *Bacillus* sp. strain and *Corynebacterium* sp. and shared certain number of CDSs with one another. On the other hand, a comparison of the gene clusters would show that strains harbor a genetic context surrounding the nit-like gene similar to that found in other strain. The percentage of similarity range among all complete amino acid sequences of nitrilase-like proteins analyzed. Nevertheless, the predicted amino

acid sequences of nitrilase-like would contain typical structural motifs. The results would highlight the potency of a bacterial strains WOD8 and WOIS2 to produce nitrilase that can be further harnessed for environmental and commercial applications. Moreover, WGS would reveal an array of nitrilase specific genes which can be effectively engineered for much enhanced production. Therefore, further studies focusing on the biochemical and structural properties of the nit protein from the two strains may also contribute to the development of sustainable bioremediation strategies.

---PAGEBREAK---

Control Number: 2022-A-97-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: A Bioinformatics Pipeline for Characterizing SARS-CoV-2 Viral Stocks

Author Block: **F. Combs**, N. Puthuveetil, A. Reese, D. Yarmosh, M. Riojas;
ATCC, Manassas, VA.

Abstract Body: The SARS-CoV-2 pandemic has highlighted the need for thorough characterization of viral stocks; because vaccine and therapeutic efficacy differ between SARS-CoV-2 variants, a well-characterized viral stock is critical for downstream research. Therefore, viral stocks must be authenticated by next generation sequencing (NGS) analysis for consensus sequence, and they must be screened for genomic variants that arise from adaptive changes due to propagation in cells. When determining the identity of an isolate in a clinical specimen, NGS reads are typically mapped to the ancestral SARS-CoV-2 sequence (AS), genomic variants are called, and a consensus sequence is generated; we refer to this sequence as the sample reference sequence (SRS). However, this process does not answer the question: has this viral stock deviated since its initial isolation and analysis? The Sequencing and Bioinformatics Center (SBC) of the American Type Culture Collection (ATCC) has developed a pipeline to answer this question by comparing NGS results of the viral stock to the AS and the SRS.The pipeline begins with NGS of the viral stock produced from early passage seed virus deposited at BEI Resources. These reads are processed to remove adapters and low-quality reads and mapped to the SRS. Then, variants are called and a consensus is generated, which we refer to as the sample consensus sequence (SCS). Finally, the SCS, SRS, and AS are aligned. This allows the identification of mutations relative to the fully annotated AS by relating positions in the SCS and SRS to their corresponding AS positions.There are five possible permutations of agreement (PoA) at each position in this alignment: 1) the SRS and SCS are identical, but the AS differs, 2) the AS and SRS are identical, but the SCS differs, 3) the AS and SCS identical, but the SRS differs, 4) all three sequences differ, and 5) all three sequences agree. Mutations of the first PoA are expected because the SRS and SCS are identical if no mutations have arisen. The second, third, and fourth PoA all indicate the potential presence of selective pressures. The second PoA signifies a mutation away from the SRS, the third PoA represents a reversion of the sample back towards the AS, and the fourth PoA suggests a new deviation. The fifth PoA covers regions of stability. With this approach, a sample that has not deviated from its initial isolation and analysis can be recognized by only having PoA of the first and fifth types. The quantity and frequency of the second, third, and fourth PoA indicate the amount of deviation that has occurred since the initial analysis. This pipeline is an important tool for quality control testing of SARS-CoV-2

variant identity and provides a means for analyzing deviation due to laboratory or natural selective pressures that ensures a solid foundation for research.

---PAGEBREAK---

Control Number: 2022-A-106-NGS

Topic 1: Microbial Chatter: Microbial ecology in health and disease

Topic 2:

Publishing Title: Validation and Implementation of a High Throughput 16S V3-V4 rDNA Sequencing Assay for Clinical Testing and Research

Author Block: **J. Williams**, H. Brochu, A. Bray, J. Crawford, L. Lyer, A. Suchanik;
Labcorp, Burlington, NC.

Abstract Body:
Validation and Implementation of a High Throughput 16S V3-V4 rDNA Sequencing Assay for Clinical Testing and Research

H. Brochu, A. Bray, J. Crawford, J. Williams L. Lyer, A. Suchanic

16S gut microbiome analysis provides physicians and patients with a non-invasive method to interrogate human gut microbiome profiles. In addition to diagnostics, Labcorp believes there is an increasing need to develop standardization and reproducibility of results for clinical utility. Automation and use of commercially available controls, standardizes the laboratory process and ensures reproducibility and quality control. The Consumer Genetics, Bioinformatics and Research and Development teams of Labcorp developed and validated an automated, high throughput 16S rDNA Gut Microbiome Sequencing Assay. This assay targets V3-V4 hypervariable regions for the relative abundance quantification of diverse microflora in the gastrointestinal tract. Fecal collection protocols were developed for at-home collection. After collection, samples are sent to the testing facility by public delivery service or Labcorp courier. DNA extraction, amplification, normalization and Illumina® sequencing are then performed using Tecan® automation, running Labcorp developed scripts. For data analysis, a custom microbiome bioinformatics pipeline was developed at Labcorp for microbiome analytics. Labcorp determined the accuracy, sensitivity, specificity, reproducibility and Limit of Detection (LoD) in the assay validation. We required a Spearman correlation > 0.8 and Bray-Curtis Dissimilarity (BCD) < 0.5 for control and donor samples to meet acceptable reproducibility standards. Multiple commercially available standards were used in validation. All standards were highly reproducible with average Spearman correlation and BCD well within the requirements of the clinical testing standards listed above. These controls were sequenced with high accuracy as calculated by their Measurement Integrity Quotients (MIQ), all of which averaged in the 90s, well above the minimum requirement of 80. Rarefaction analysis of controls assessed the sensitivity and specificity. The commercially available controls were found to have sensitivities and specificities within target range, with multiple sensitivities of 100% and all specificities over 99%. Labcorp assessed the intra- and inter-assay reproducibility and the LoD of our assay using donor fecal swabs. All intra- and inter-assay replicates met the reproducibility standards, provided they satisfied the LoD genus mapping threshold. In summary, Labcorp believes there is a large undetermined role for Microbiome testing in clinical diagnostics. Applying clinical standards and validation to a high-throughput, highly reproducible 16S Gut Microbiome assay, enables Labcorp to support the rapidly growing field of microbiome research and future clinical testing.

---PAGEBREAK---

Control Number: 2022-A-108-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Using Next Generation Sequences (NGS) to advance Identity (ID) Testing in the Biosafety Testing Field

Author Block: **R. A. Bova**;
MilliporeSigma, Rockville, MD.

Abstract Body: The use of Next Generation Sequencing (NGS) in biosafety testing has grown significantly over the past decade. With new advancements in technology and development of methods there has never been a better time to integrate NGS into regulated testing plans. Because regulators have set high expectations on drug manufacturers to establish the sequence identity and purity of their viral and non-viral vectors, it is important to have a validated method to ensure reliable sequence confirmation along with variant detection. Not having the correct sequence could have unintended consequences resulting in delays and loss of time and money for valuable products needed on the market. For this reason, it is recommended that NGS be used in the early parts of development testing to confirm starting materials such as the plasmid or virus seed stocks, as well as in the production stage to establish vector identity and production consistency. Before the advent of NGS in the field, the previous standard for identity confirmation was using various molecular tests, the most common being Sanger sequencing. Although still applicable for establishing a majority consensus sequence, Sanger has its limitations such as inability to detect low frequency variants, inability to sequence distal ends of the molecule and fragmented coverage of the molecule of interest due to primer placement. By using NGS, these obstacles are able to be overcome. In addition, complex regions can be easily sequenced, for example the inverted terminal repeat (ITR) region commonly found in adeno-associated viruses (AAV), and extraordinary depth of coverage can be achieved. The benefits of using NGS technology over Sanger sequencing are shown in a recent case study that will be presented involving the analysis of an AAV. The NGS method and subsequent analysis is easily able to identify a subtle sequence variation whereas Sanger sequencing misses this, showing how NGS provides higher confidence for the sequence confirmation of the product being assessed that past technologies have not allowed.

---PAGEBREAK---

Control Number: 2022-A-113-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Combining Mash and NCBI Datasets via Nextflow for a *Legionella* Species Identification Tool

Author Block: **J. Hamlin**, M. Willby, J. Winchell;
Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases, Respiratory Diseases Branch, Atlanta, GA.

Abstract Body: Nationally, Legionnaires' disease (LD) cases have been increasing since 2000, but current laboratory identification of species is limited. *Legionella pneumophila* is the most

prevalent *Legionella* species of clinical concern in the United States; however, other *Legionella* species are associated with disease. Clinical diagnosis of LD commonly occurs via the urinary antigen test which only detects *L. pneumophila* or multiplex real-time PCR, with targets that may not distinguish between all *Legionella* species. Thus, with either test, species level identification of a non-pneumophila *Legionella* is not possible. We developed a tool for *Legionella* species identification. Our species identification tool uses the program Mash to evaluate the similarity between an isolate of interest and a database generated from all *Legionella* species. Options include provision of a pre-built Mash database or creation of a database each time the tool is used. We implement database creation using a text file containing the organisms of interest and access the NCBI genome database using the NCBI Datasets tool to download all available genomes corresponding to those organisms. The input for the tool is next-generation sequencing reads derived from a *Legionella* isolate (gzipped fastq files). The identification of the isolate occurs via comparison against a Mash database. The output is a plain text file listing the top five species matches. One argument of interest a user can specify is the max_dist flag, which tests against the values of the Mash distance measurement modulating the specificity of reported matches. For *Legionella*, we estimate a conservative Mash distance of 0.05 between species. If no result has a value less than or equal to this, then the best species match is not indicated though the top five matches are still output. Additionally, we use Nextflow to wrap our pipeline to provide portability and reproducibility. We are evaluating the accuracy and reproducibility of the tool using a dataset focusing on *Legionella* species to meet the requirements for laboratory-developed tests certified under Clinical Laboratory Improvement Amendments (CLIA) regulations. The tool is currently available on GitHub: https://github.com/jennahamlin/mashwrapper, which includes a test dataset of five *Legionella* isolates. The tool is flexible enough to work with any organism that can be downloaded from NCBI and run through Mash.

---PAGEBREAK---

Control Number: 2022-A-120-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Species-Specific Subgrouping and Whole Genome-Based Identification of *Pseudomonas aeruginosa* Rectifies Taxonomy Assignment within the NCBI Database

Author Block: **S. Park**[1], N. A. Hasan[2], J. Chun[1];
[1]CJ Bioscience Inc., Seoul, KOREA, REPUBLIC OF, [2]EzBiome Inc., Gaithersburg, MD.

Abstract Body: **Background** Identification of bacterial isolates relies upon various methods such as matrix-assisted laser desorption/ionization time-of-flight spectrometry. Whole genome-based identification of bacterial isolates with average nucleotide identity (ANI) uses standardized NGS methodology with set numerical parameters for species delimitation. However, sometimes isolates are misidentified or fail to be assigned to the correct species due to a lack of coverage in reference databases, yet are still found in public resources. Even if the species-level identification is accurate, higher resolution below the species-level, which is important especially for clinically relevant bacterial species, is limited. Here, we devised a workflow for bacterial identification at species-level based on whole genome sequencing data, followed by species-specific subgrouping using a single nucleotide variation (SNV)-based core genome method. To verify our method, we focused on *Pseudomonas*

aeruginosa, an important opportunistic pathogen from a highly diverse genus that have been isolated from a wide range of environmental and clinical sources, and represented their population structure. **Methods** Genome sequences from the *Pseudomonas* genus were downloaded from the EzBioCloud genome database. Each genome was identified at species-level by calculating OrthoANI with an ANI cutoff of 95%. The genomes were then grouped by the number of SNVs relative to their core genome size, using a 1% threshold. **Results** A total of 10,514 *Pseudomonas* genomes including 5,445 *P. aeruginosa* were identified using ANI. By comparing these respective genome taxon names with those in NCBI, we found that some *P. aeruginosa* genomes were misidentified, such as *Pseudomonas* sp. P179. Also, genomospecies CP000744_s genomes were identified as *P. aeruginosa* despite 93.6~93.8% ANI values against the type strain. Additionally, as a result of the species-specific clustering, 10 subgroups of *P. aeruginosa* were obtained. These subgroups not only form two major clades which are consistent with previous research but also show correlations among the strains isolated from different studies. **Conclusion** Our workflow demonstrates that an accurate and consistent identification method using whole genome sequences is necessary for the proper identification of bacteria since it would prevent erroneous diagnoses based on the misidentification of targets. Furthermore, since species-specific subgroups are identified by comparing core genome SNVs, this will allow for newly isolated strains to be easily identified without the need for de-novo phylogenetic tree reconstruction.

---PAGEBREAK---

Control Number: 2022-A-127-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Use of recombinant bacteria with unique tags as spike-in controls for the quantification of microbiome content

Author Block: **L. Papazisi**[1], R. Chuang[1], B. Tang[1], M. Hunter[2], S. J. Green[3], J. Lopera[2], B. Benton[2]; [1]American Type Culture Collection (ATCC), Gaithersburg, MD, [2]American Type Culture Collection (ATCC), Manassas, VA, [3]University of Illinois at Chicago, Chicago, IL.

Abstract Body: Advanced sequencing and bioinformatics technologies have revolutionized microbiome research in remarkable ways, opening up applications in diagnostics, therapeutics, and environmental sciences. Despite the promise of these technologies, the analysis of metagenomic data remains challenging due to the technical biases introduced throughout the metagenomics workflow—from sample preparation to bioinformatic analysis. Further, the natural complexity of microbial communities themselves has challenged microbiome researchers in their ability to make meaningful, quantifiable, reproducible, and comparable measurements across different laboratories. To help promote assay standardization and validation, ATCC has developed innovative spike-in standards for microbiome research. These controls are prepared as whole cell or nucleic acid mixtures comprising three genetically engineered bacterial strains (derived from Escherichia coli, Staphylococcus aureus, and Clostridium perfringens), each containing a unique synthetic DNA tag that can be detected and quantified in routine 16S rRNA gene amplicon and shotgun sequencing assays. To demonstrate the utility of these spike-in controls in microbiome research, we conducted studies where we mixed them with whole-cell or gDNA mock communities containing different bacterial strains at various ratios. The resulting data showed that the unique tags of all three bacteria were identifiable and quantifiable by shotgun and 16S rRNA

amplicon sequencing. These proof-of-concept experiments support the utility of using spike-in controls with a unique 16S rRNA tag to monitor the full process from DNA extraction to data analysis of a microbiome workflow for both 16S rRNA and shotgun metagenomics assays.

---PAGEBREAK---

Control Number: 2022-A-128-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Microbial cooperation and competition in polyethylene terephthalate degrading consortia

Author Block: **L. Schaerer**, R. Wu, L. Putman, R. Ong, S. Techtmann;
Michigan Technological University, Houghton, MI.

Abstract Body: Polyethylene terephthalate is a popular, inexpensive, and durable plastic which is used for packaging, disposable water bottles, and synthetic fabrics. Over 70% of plastic waste currently pollutes landfills and oceans because current plastic recycling methods are inadequate. Polyethylene terephthalate degrading bacteria have been isolated from numerous environments, including compost. These organisms completely degrade polyethylene terephthalate by using carbon from plastic to support their growth. However, biological rates of polyethylene terephthalate depolymerization are too slow to efficiently degrade plastic in industrial settings. Chemical depolymerization is an alternative which eliminates the need for biological depolymerization and yields aromatic monomers which are more quickly biodegraded. Here we use two terephthalate degrading bacterial consortia enriched from compost to explore cooperation and competition in bacterial communities. The consortia were grown on the mixture of compounds resulting from chemical depolymerization of polyethylene terephthalate (terephthalamide, terephthalate, ethylene glycol, and a mixture of all three products). We extracted DNA and RNA from the cultures, and performed 16S rRNA, metagenomic, and metatranscriptomic sequencing. Short-read metagenomic sequencing of one consortium yielded 18 high quality metagenomic bins, representing the phyla Proteobacteria, Actinobacteria, and Acidobacteria and aromatic-degrading genera *Brevundimonas*, *Achromobacter*, *Hydrogenophaga*, and *Rhodobacter*. Using long-read metagenomic sequencing, we recovered 30 circular, high quality metagenomic bins representing the phyla Proteobacteria and Actinobacteria, including the genera *Brevundimonas*, *Hyphomonas*, *Hydrogenophaga*, and *Pelagibacterium*. To understand how microbial communities cooperate or compete to degrade polyethylene terephthalate, we employed metatranscriptomics and flux balance analysis to explore each microorganism's expression of aromatic-degrading genes, and to detail each organism's role in the consortia. Long-read sequencing data allowed for more complete genomes from the consortia and better reconstruction of division of labor within the microbial community. By studying chemically depolymerized polyethylene terephthalate degrading microbial communities, we lay the foundation for future industrial systems that couple chemical and biological processing to degrade plastic waste.

---PAGEBREAK---

Control Number: 2022-A-129-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Development and Evaluation of Bioinformatic Approach to Predict Antibiotic Susceptibility of *Mycobacterium tuberculosis*

Author Block: M. Ezewudo, L. Cowan, **J. Posey**;
1Division of Tuberculosis Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA.

Abstract Body: In 2018, the National Tuberculosis Molecular Surveillance Center (NTMSC) began whole genome sequencing for at least one isolate for every U.S. culture-confirmed tuberculosis case. Data are analyzed for determination of whole genome multi locus sequence types and nomenclature assignment. Results are shared with public health partners through the web-based TB Genotyping Information Management System. We developed a bioinformatic approach to utilize whole genome sequence data for the prediction of antibiotic susceptibility of samples to the first line drugs rifampicin, isoniazid, ethambutol and pyrazinamide by analyzing genetic loci associated with resistance. We used open-source tools for sequence reads alignment (BWA), coverage analysis (Samtools), variant calling (GATK-Mutect2) and annotation (SnpEff). Mutect2 was chosen as it provided accurate identification of low frequency phased alleles. For interpretation, we developed a database of variants that harmonized data from different sources including a WHO mutation catalog, data from CDC's Molecular Detection of Drug Resistance Service, and from homoplasy analysis of U.S. data from the NTMSC. All variants were initially classified as resistance unknown and modified to susceptible or resistant following comparison to the database. Loss of function variants in selected loci were also interpreted as resistant. Samples lacking variants in selected loci were considered susceptible to first-line antibiotics. The pipeline was evaluated on a set of 14,922 samples from U.S. TB cases between 2018-2020 with reported phenotypic drug susceptibility test data. After excluding variants with unknown association with resistance, the sensitivity of predicted resistance to isoniazid was 93%, rifampicin was 96%, pyrazinamide was 52% and ethambutol was 86%. The sensitivity of predicted resistance to pyrazinamide varied by lineage with 6% for lineage 1 and 60-78% for lineages 2,3 and 4. Predicted results were also compared for a subset of samples with susceptibility test data for all four first-line drugs (n=13,259). The accuracy of predicting susceptibility to all four drugs was 97%. The accuracy of susceptibility prediction varied by lineage and was 90% for lineage 1 and 99% for lineage 2,3, and 4. In a low drug resistance setting, rapid and accurate prediction of susceptibility has the potential to alter testing algorithms and reduce the need for slower phenotypic drug susceptibility testing.

---PAGEBREAK---

Control Number: 2022-A-141-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Long read genome assemblers are prone to missing small plasmids

Author Block: **J. Johnson**, K. Jacob, M. Soehnlen, H. Blankenship;
Michigan Department of Health and Human Services, Bureau of Laboratories, Lansing, MI.

Abstract Body:

Hybrid genome assembly, using long and short read DNA sequencing, has greatly improved our ability to resolve bacterial genomes. This approach is generally split into two categories: short-read-first and long-read-first assembly methods. While long-read-first hybrid assembly has demonstrated advantages over short-read-first assembly, there is some evidence that long-read assemblers may fail to recover small plasmids (< 10 kb). This work aimed to investigate the ability of common long read assemblers (i.e., Flye, Raven, and Miniasm) to recover small plasmids during de novo bacterial assembly. This was accomplished by determining the number of times each assembler correctly recovered all contigs present in completed bacterial assemblies from both simulated and lab generated, long and short reads datasets. These results were compared to the performance of the short-read-first assembler, Unicycler. For the simulated datasets, ONT long reads and Illumina short reads were generated from four complete FDA-ARGOS *Escherichia coli* genomes using the Badread and Art read simulators. Contigs from these completed assemblies were then searched in the new assemblies using blastn. 3 *E. coli*, 3 *Klebsiella pneumoniae*, and 2 *Neisseria gonorrhoeae* isolates were then sequenced using an ONT Mk1C and Illumina MiSeq and complete genome assemblies created for each isolate from all individual assemblies using the Trycycler method. These completed assemblies were used as references to determine the number of times each individual assembly correctly recovered each contig. Simulated datasets predicted the complete loss of contigs < 10 kb by all long read assemblers and only 38% recovery of these contigs using Unicycler. By contrast, Flye, Raven, and Miniasm exhibited recovery rates of 15, 33, and 39 % for < 10 kb contigs when using lab generated reads, while Unicycler had a recovery rate of 100%. These results demonstrate that long read assemblers are prone to missing plasmids smaller than 10 kb.

---PAGEBREAK---

Control Number: 2022-A-144-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Genome mining for bioactive secondary metabolites in surfactin producing *Bacillus subtilis* strain E2-03

Author Block:

S. O. Akintayo[1], B. Neumann[2], M. Vahidinasab[1], M. Henkel[1], L. Lilge[1], R. Hausmann[1];
[1]University of Hohenheim, Department of Bioprocess Engineering, Stuttgart, GERMANY, [2]Institute for Hospital Hygiene, Medical Microbiology and Clinical Infectiology, Paracelsus Medical University, Nuremberg General Hospital, Nuremberg, GERMANY.

Abstract Body:

Based on their biosynthetic and metabolic potentials, *Bacillus* strains are used for numerous applications in biotechnological, food, and agricultural fields. Strain E2-03 isolated from palm oil mill effluent, in the process of screening for lipopeptide-producing *Bacillus* strains from food sources, was confirmed to produce surfactin. Whole genome sequencing (WGS) revealed a genome size of 4.8 mb with a GC content of 43.5% and 4318 protein-coding genes. Species identification was realized using the raw reads as well as the reconstructed sequences with the Type Strain Genome Server (TYGS) online platform. Strain E2-03 was identified as *Bacillus subtilis* subsp. *subtilis*, and one COL (SD853)-type plasmid with a size of 9 kb was identified. E2-03 harbors the phosphotransferase gene *mphK*, mediating resistance to macrolide antibiotics, and the gene *aadK*, encoding an aminoglycoside 6-adenylyltransferase, mediating resistance to aminoglycosides. Genome-scale metabolic model (GEM) was constructed using KBase implemented Build Metabolic Model App based

on the ModelSEED Pipeline for genomes. GEM predicted 1336 reactions and 1340 compounds involving 1022 genes in E2-03. The genome mining pipelines antiSMASH 6.0 and BAGEL 4 were used to identify genes and gene clusters encoding bioactive secondary metabolites and bacteriocin (including RiPP), respectively. Altogether, strain E2-03 uses 8.9% of its genome for the synthesis of secondary metabolites, with a total of 12 biosynthetic gene clusters (BGCs) detected, including six non-ribosomal peptide synthetases (NRPSs), one polyketide synthase (PKS), two bacteriocins, two terpenes and one tRNA-dependent cyclodipeptide synthase (CDP). NRPSs of lipopeptides surfactin and fengycin were identified in addition to other secondary metabolites such as bacilysin, subtilosin A, and bacillibactin. Further investigation into the modular organization and functionality of the NRPSs, as well as structural analysis of synthesized non-ribosomal peptides, alongside their applicability for biotechnological, food, and agricultural purposes would be carried out.

---PAGEBREAK---

Control Number: 2022-A-149-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Maximizing MAGs from long-read metagenomic assemblies: a new post-assembly pipeline with circular-aware binning

Author Block: **D. Portik**, J. Wilkinson;
PacBio, Menlo Park, CA.

Abstract Body: There are many challenges involved with metagenome assembly, including the presence of multiple species, uneven species abundances, and conserved genomic regions that are shared across species. Highly accurate long reads can overcome many of the obstacles associated with metagenome assembly. PacBio HiFi sequencing of metagenomic samples with the Sequel IIe system regularly produces reads 8-15kb in size with a median QV ranging from 30-45 (99.9-99.99% accuracy). With the development of new metagenome assembly algorithms specific to HiFi reads (including hifiasm-meta), it is now possible to reconstruct full metagenome-assembled genomes (MAGs) for many high abundance species. However, discontiguous assemblies still occur for lower abundance taxa, and post-assembly tools are required to identify MAGs in this category. Here, we present the HiFi-MAG-Pipeline, a comprehensive workflow for processing long-read metagenome assemblies. This open-source pipeline is written in a workflow management language (snakemake) and automates all major steps including binning, quality filtering, and taxonomic identification. The outputs include high-quality MAG sequences, associated metadata, and visualizations of key MAG characteristics. Importantly, we found that typical binning strategies can exhibit unexpected behavior when the assembly contig set contains complete, circular contigs. In these cases, mis-binning inflates contamination scores for these contigs and causes them to be eliminated from the final MAG set. To address this issue, we developed and implemented a new circular-aware binning strategy in our workflow. To demonstrate the use of our post-assembly pipeline, we analyzed several publicly available HiFi metagenomic datasets. We assembled each dataset using hifiasm-meta and then used the HiFi-MAG-Pipeline to process the assemblies. We also performed a direct comparison between our circular-aware binning strategy and a standard binning method. Using circular-aware binning, we found a 7-16% increase in total high-quality MAGs and a 16-66% increase in the number of single contig MAGs recovered. Furthermore, we found that 35-68% of the high-quality MAGs across

samples were composed of a single contig. The latter result demonstrates hifiasm-meta routinely assembles complete genomes from HiFi metagenomic datasets. Finally, we show that for a given sample there is a predictable relationship between the total amount of sequencing data and total number of MAGs recovered. Together, our work demonstrates 1) HiFi reads can be used to produce high-quality metagenome assemblies, and 2) workflows specific to long-read assemblies should be used to maximize MAG recovery.

---PAGEBREAK---

Control Number: 2022-A-154-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Investigation of Changes in the Resistome and Phageome of Lab-scale Moving Bed Biofilm (MBB) and Suspended Sludge (SS) Reactors under Tetracycline Exposure

Author Block: **K. Yanac**, Q. Yuan, M. Uyaguari;
University of Manitoba, Winnipeg, MB, CANADA.

Abstract Body: Wastewater treatment plants (WWTPs) and natural water bodies receiving treated wastewater have been considered as hotspots for antibiotic resistance. Richness and abundance of the microbiome and environmental and operational factors in WWTPs promote antibiotic resistance development and mobilization by several mechanisms, including horizontal gene transfer (HGT) and vertical gene transfer (VGT). Since HGT can occur between different bacterial species, antibiotic resistance genes (ARGs) can be transferred to the various environmental and pathogenic bacteria. While the role of plasmids and transposons (conjugation) in the spread of ARGs in environmental settings has been relatively well investigated, the role of bacteriophages (transduction) still waits to be elucidated. The literature on the role of bacteriophages in the dissemination and acquisition of antibiotic resistance is very limited and controversial. Considering the current findings, this study aims to provide insights into the role of bacteriophages as potential carriers of ARGs and investigate changes in resistome and bacteriophage profiles in MBB and SS reactors under tetracycline exposure using metagenomics. To achieve these objectives, one advanced and one conventional lab-scale wastewater treatment system, employing MBB and SS, respectively, were fed with synthetic wastewater over 6 months. Tetracycline concentration in the feed was increased every two months (0, 100 and 300 μg/L, respectively). Sludge, effluent and biofilm samples were collected from reactors every two months and processed to separate and concentrate bacterial and bacteriophage fractions using physical and chemical techniques. Extracted DNA samples were sequenced on NovaSeq6000 S4 paired-end 150. Reads were assembled using metaSPAdes after quality check and trimming. Contigs larger than 1000 bp were processed with VirSorter, VirSorter2, VIBRANT and virfinder to identify viral contigs.. Bacterial and Phage contigs were binned using VAMB and vRhyme, respectively. Completeness of bacterial and phage meta genome-assembled genomes (MAGs) were assessed using checkM and checkV, respectively. Phage and bacterial contigs were annotated using Prokka. Annotated genes were searched for ARGs using BLASTp against comprehensive antibiotic resistance database ((CARD). Taxonomical classification of bacterial and phage MAGs were assessed using GTDB-tk and vConTACT2, respectively. Preliminary results indicated that bacteriophages might promote the spread and acquisition of ARGs. We expect to assess ARG-carrying phages, the prevalence of ARGs in bacterial and phage fractions, and phage-host interactions under

different tetracycline concentrations. We are still analyzing the data and we are expecting to have more in-depth results in few months.

---PAGEBREAK---

Control Number: 2022-A-157-NGS

Topic 1: Microbial Chatter: Microbial ecology in health and disease

Topic 2:

Publishing Title: Metagenomic insight into Chrysophyllum albidum spoilage implicates respective decline and increase in Pseudomonas and Acetobacter abundance

Author Block: **V. Ezebuiro**, A. E. Ataga, N. G. Ogbuji, U. C. Nwosu;
University of Port Harcourt, Port Harcourt, NIGERIA.

Abstract Body: This study investigated the spoilage of African star apple (*Chrysophyllum albidum*) by bacteria using metagenomic approach. Healthy and diseased samples of *Chrysophyllum albidum* were obtained from Choba market in Port Harcourt, Rivers State, Nigeria. Bacterial DNA was extracted from both healthy and diseased samples and subjected to 16S rDNA sequencing and metagenomic analyses. Metagenomic analyses of bacterial strains from the samples revealed total operational taxonomical units (OTUs) from healthy and diseased samples as 113 and 228, respectively. Proteobacteria were dominant in both the healthy and diseased samples. For healthy African star apple, the most abundant genus was *Pseudomonas,* representing about 65% of the OTUs followed by *Candidatus Portiera.* For diseased *Chrysophyllum albidum* sample, *Pseudomonas* was the most abundant (with relative abundance of approximately 85%) genus, followed by *Acetobacter*. Comparative analyses of the relative abundance of the individual genera revealed significant reduction in the abundance of *Pseudomonas* in the diseased samples compared to healthy samples. Conversely, relative abundance of *Acetobacter* genus increased in the diseased *Chrysophyllum albidum* samples compared to the healthy samples. Additionally, *Candidatus Portiera,* which was present in healthy samples was not found in the diseased samples. The current study has helped in recognizing the microbial community structure of healthy and diseased samples of *Chrysophyllum albidum*. These findings can help predict bacterial community structural dynamics involved in the spoilage of African star apple (*Chrysophyllum albidum*) and thus how the spoilage can be prevented or controlled.

---PAGEBREAK---

Control Number: 2022-A-158-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Characterization of Mock Fungal Organisms by 3 Bioinformatics Pipelines

Author Block: **N. Mok**[1], G. Van Domselaar[2], A. Bar-Or[3], C. N. Bernstein[1], J. D. Forbes[4], M. Grahm[2], C. Bonner[2], J. Hart[5], R. A. Marrie[1], J. O'Mahony[6], E. A. Yeh[4], F. Zhu[7], M. Bakker[1], B. Banwell[8], E. Waubant[5], H. Tremlett[7], N. Knox[2];
[1]University of Manitoba, Winnipeg, MB, CANADA, [2]Public Health Agency of Canada, Winnipeg, MB, CANADA, [3]University of Pennsylvania, Philadelphia, PA, [4]University of Toronto, Toronto, ON, CANADA, [5]University of California San Francisco, San Francisco,

CA, [6]Hospital for Sick Children, Toronto, ON, CANADA, [7]University of British Columbia, Vancouver, BC, CANADA, [8]Perelman School of Medicine University of Pennsylvania, Philadelphia, PA.

Abstract Body:

Background: The gut microbiota is thought to play a role in immune-mediated diseases such as multiple sclerosis (MS). Recent studies have associated MS with the gut mycobiota—the fungal component of the microbiome. However, standardized methods for evaluating the gut mycobiota are lacking and require systematic evaluation of sequencing protocols, reference databases, and bioinformatics pipelines to judge their suitability for investing in the gut mycobiome.

Objectives: We evaluated the performance of different sequencing approaches by varying the PhiX concentration. We also assessed the abilities of different analytical pipelines to characterize a pre-defined fungal community.

Methods: Using a mock-community control of 19 defined fungal organisms, we targeted the fungal internal transcribed spacer 2 (ITS2) region using the Illumina MiSeq platform. Concentrations of 25% and 50% of PhiX control were tested in technical replicates to address low-diversity amplicon sequencing. Generated sequences were characterized using the UNITE database—a curated collection of eukaryotic ITS sequences—in conjunction with three fungal sequence analysis pipelines: LotuS v1, mothur, and PIPITS.

Results: A 50% PhiX spike-in increased the sequence quality but decreased the overall number of reads by more than half when compared to a 25% spike-in (~500,000 target reads at 50% PhiX spike-in vs. ~2 million target reads at 25% PhiX spike-in). An assessment of the UNITE v8.2 database indicated that one of 19 mock-community organisms was absent from the database and could not be identified in any analysis. We also found inconsistent use of the fungal taxonomic nomenclature applied within the UNITE database such as telomorphic vs. anamorphic naming conventions. Out of 19 mock-community organisms, the LotuS pipeline correctly classified 6 species, whereas mothur correctly identified 5 species, and PIPITS correctly identified only 3 species.

Conclusions: The LotuS pipeline identified the most fungal species in the mock-community. The sequence read quality was optimal when 50% PhiX spike-in was used for sequencing ITS2 amplicon libraries, but the overall number of reads decreased substantially. The UNITE database is unable to fully characterize all 19 members of our mock-community. Validation of an amplicon-based sequencing and analytical approach for mycobiome characterization using a mock-community with known fungal identities will aid in the accurate characterization of the mycobiota in future studies.

---PAGEBREAK---

Control Number: 2022-A-159-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Using Microbial Communities and Machine Learning to Detect Oil Contamination in the Great Lakes

Author Block:

I. Bigcraft, A. Kuntzleman, E. Byrne, S. Techtmann;
Michigan Technological University, Houghton, MI.

Abstract Body:

Traditional data analysis techniques for data produced by NGS methods often fail to comprehensively analyze the full range of information present in the data. Machine learning methods offer the potential to more fully appreciate the depth of contextual information

present in 'omics datasets. Although employing machine learning is not necessary to sufficiently answer many biological hypotheses, some problems lend themselves to machine learning approaches. Biosensing, using microbial indicators to detect environmental states or changes, is one such problem. While many approaches have sought to identify single indicator species in datasets to serve as biomarkers, the complexity of natural microbial communities and the interactions therein have complicated biomarker selection for environmental monitoring. Accurately identifying such biomarkers requires considering contextual information from across the dataset; a task well suited for machine learning.In this study we applied machine learning to the biosensing problem of detecting oil in samples from the Great Lakes. A large-scale oil spill in the Great Lakes could be catastrophic for the local ecosystem. While spotting such a spill after it has happened would obviously not require specialized biosensing, detecting transient oil spills and trace amounts of oil could be used to identify potential spill risks. Here we used 16S rRNA sequencing from 3 different Great Lakes oil amendment microcosm studies, totaling to 499 samples across seven locations and three seasons, to develop a model to predict the presence of oil based on community composition. Although many other study designs are limited to training within a single uniformly sampled study, our varied dataset allows investigating the generalizability of our models to different locations and timepoints. Our preliminary results indicate strongly signaled changes to the microbial community in response to oil: a humble Random Forest model attained over 90% accuracy on held out data when trained and tested on data from a single study. However, when tested on data from different seasons, locations in the Great Lakes, and sequencing runs, model accuracy fell to a less satisfactory 50-75%, depending on which of the three studies were used. In the context of a consistent sample space, the strategy of testing on held out data is generally sufficient to show generalizability. However, in the context of 16S rRNA community sequencing regional, seasonal, and temporal differences in community composition can substantially alter a machine learning model's performance and generalizability. Therefore, consideration for the generalizability of machine learning models is critical when working with complex biological datasets. Utilizing data from a wider range of spatial and temporal conditions may be necessary to identify and address potential problems with model generalizability.

---PAGEBREAK---

Control Number:	2022-A-165-NGS
Topic 1:	Microbial Chatter: Microbial ecology in health and disease
Topic 2:	
Publishing Title:	Metagenomic Analysis Reveals Seasonal Patterns for DNA and RNA viruses in Urban Aquatic Environments Influenced by Anthropogenic Activities.
Author Block:	**J. D. Francis**, M. Uyaguari; University of Manitoba, Winnipeg, MB, CANADA.
Abstract Body:	“Pure effluents” discharged into the Red and Assiniboine Rivers of Winnipeg may not be fully treated as traditional methods that monitor the microbial quality of wastewater focus solely on the detection of fecal indicator bacteria. The lack of virome analyses in aquatic ecosystems creates an opportunity for our study to explore viral communities in urban settings. We hypothesize that viral community structures from wastewater may remain unchanged in urban influenced environments. Our study aims to characterize viral DNA and RNA community structures present in the Red and Assiniboine rivers using metagenomic

analysis.
To evaluate seasonal variabilities of viruses, aquatic sampling was conducted at 11 locations along both rivers during spring, summer and fall 2021. One sample was collected from a lesser polluted environment to represent a case-control study and another consisted of MilliQ water for a background control. Samples were filtered and underwent skimmed milk flocculation for viral concentration. Total nucleic acids were extracted, separated into half and enriched enzymatically for viral DNA and RNA. 15 µL of viral DNA and randomly amplified viral RNA were sent for NGS to generate ~4 M Pair Ended reads and 2.4 Gb/sample sequence. A mock community of pooled DNA and RNA viruses was included to account for metagenomic sequencing controls. Kraken 2 viral genome database was used to identify assembled reads as DNA and RNA viruses.
The most abundant (%) DNA viruses that were consistently identified in all seasons from the Kraken 2 viral genome database were: *Myoviridae* (15-30%), *Podoviridae* (9-17%) and *Siphoviridae* (30-55%). The most abundant (%) RNA viruses identified were: *Retroviridae* (36-97%) in spring 2021, *Virgaviridae* (64-88%) and *Partiviridae* (7-25%) in summer 2021 and *Virgaviridae* (5-71%), *Retroviridae* (9-65%) and *Picobirnaviridae* (9-26%) in fall 2021. No viruses were identified in MilliQ samples with lower percentages of the previously mentioned viruses identified in the lesser polluted environment. These findings suggest that DNA viruses were relatively more stable and consistent throughout changes in seasonal weather while RNA viruses were seasonally more variable possibly owing to their genomic instability. Whereas unclassified DNA and RNA viromes were present in all aquatic samples, a higher percentage of RNA viromes remained unclassified possibly due to the small size of RNA viruses that results in low yields and is prone to degradation.
Previous research conducted in the Uyaguari lab evidenced human enteric viruses such as *Adenovirus* in raw sewage samples collected from Winnipeg's sewage plants. Although DNA viruses were found to be more stable than RNA viruses, the identification of similar viruses in "pure effluent" samples indicates the lack of effective viral treatment practices at sewage plants and thus confirmed our hypothesis to be true. Future directions may involve additional screening over longer sampling periods.

---PAGEBREAK---

Control Number: 2022-A-168-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Does Microbial Dark Matter matter?

Author Block: **H. Barak**, N. Fuchs, A. Sivan, A. Kushmaro;
Ben Gurion University, Beer Sheva, ISRAEL.

Abstract Body: Microorganisms are the most diverse and abundant life forms on Earth. Nevertheless, in many environments more than 99% of them remain uncultured. To date, our knowledge regarding microbial life is still lacking, as it is based mainly on information gleaned from cultivated microorganisms. Microbiologists termed this 'uncultured microbial majority' as 'microbial dark matter' (MDM), a term borrowed from astrophysics. The realization of how diverse and unexplored microorganisms are, actually stems from recent advances in molecular biology, and in particular from the sequencing of the microbial small subunit ribosomal RNA gene directly from environmental samples. Using next-generation

sequencing (NGS) analysis (both 16S rRNA and whole genome shotgun), sequences are compared to reference databases that contain only a small part of the existing microorganisms. Therefore, their taxonomy assignment reveals groups of unknown microorganisms. Interestingly, despite their great importance these unknowns are usually ignored and rarefied from diversity analysis. In the present work, we analyzed the 16S rRNA gene sequences of microbial communities found in four different environments- a living organism, a desert environment, a natural aquatic environment, and a membrane wastewater reactor. From those datasets, we chose sequences of potentially unknown bacteria and validated their existence by specific amplification and re-sequencing. Sequences were compared to different databases such as NCBI BLAST, RDP, and Silva. Representative unknown sequences were integrated into the tree of life, revealing potentially new candidate phyla (and other lineages). These microbial dark matter sequences (MDMS) were also screened against Metagenome-Assembled Genomes (MAGs) from the explored environments for additional validation as well as for taxonomic and metabolic capacity characterization. This study provides further evidence for the great importance of microbial dark matter sequences (MDMS) in environmental metataxonomic analysis of 16S rRNA gene and in the discovery of novel organisms.

---PAGEBREAK---

Control Number: 2022-A-171-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Assemblies Matter: AMR Detection

Author Block: **E. Litrup**, K. Loaiza;
Statens Serum Institut, Copenhagen, DENMARK.

Abstract Body: **Introduction** Detection and analysis of antimicrobial resistance (AMR) genes and point mutations (PM) in genomes of microbes are vital. Both different analysis and databases of AMR genes and PMs are available to provide results; however, often not immediately comparable due to differences in nomenclature, level of curation and input data. Similarly, input data, in the form of different assemblies using identical analysis workflow can also change the output. Here, we analyzed the genomes of strains isolated from humans with infections caused by *Salmonella*, Shiga toxin-producing *E. coli* (STEC) or *C. difficile*. Two different assembly methods was compared by using AMRFinderPlus. Detected AMR targets using the different assemblies were also compared.

Material and Methods All strains were sampled as part of the Danish human surveillance and sequenced at SSI. Strains containing known AMR genes were selected. 500 genomes of *Salmonella*, STEC and *C. difficile* respectively were analyzed. SPAdes assemblies were run using default settings in BioNumerics8.1 (Applied Maths) and SKESA 2.4.0 assemblies were run with default settings in Bifrost (https://github.com/ssi-dk/bifrost). Assembly statistics were calculated for both methods and assemblies were run through AMRFinderPlus 3.10.30 (db version 2022-05-26.1) with custom settings (--ident_min 0.9 --coverage_min 0.9). Detected AMR core genes and resistance associated PM's were enumerated and visualized using custom script written in Python 3.10.5.

Results The analysis of the different assemblies showed that for more than 95% of isolates in all three species SPAdes produced larger genomes, longer contigs and a higher N50 and

N90. SPAdes also produced fewer contigs in more than 90% of isolates in all three species. For *C. difficile*, 8 differences were detected in 4 classes of antimicrobials and differences were detected in 7 strains (1.4% of total dataset). The differences included 4 genes (in 2 classes) only detected by SPAdes and 4 other genes (in 2 classes) only detected by SKESA. For STEC, 25 differences were detected in 7 classes of antimicrobials and the differences were detected in 21 strains (4.2% of total dataset). The differences included 16 genes (in 5 classes) only detected by SPAdes and 9 other genes (in 2 classes) only detected by SKESA. For *Salmonella*, 31 differences were detected in 8 classes of antimicrobials and the differences were detected in 26 strains (5.2% of total dataset). The differences included 20 genes (in 8 classes) only detected by SPAdes and 11 other genes (in 4 classes) only detected by SKESA. **Conclusion** SPAdes performed better in the case of AMRFinderPlus: more genes were detected however, SKESA assemblies contained genes not detected in the SPAdes assemblies.

---PAGEBREAK---

Control Number: 2022-A-174-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: An End to End Pipeline for Characterization and Annotation of Traceable Bacterial Material

Author Block: **J. Bagnoli**, D. Yarmosh, N. Puthuveetil, F. Combs, A. Reese;
ATCC, Gaithersburg, MD.

Abstract Body:
The need for well characterized quality genomics data is crucial for life science research. Laboratories often leverage publicly available data as a cornerstone for their experimental design. While such databases have grown exponentially via contributions from the scientific community, data provenance is often lacking, and authenticity of the underlying materials is not assured.

American Type Culture Collection (ATCC) has discussed this issue before (Yarmosh, D. A. et al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. mSphere 7, e00077-22 (2022) and has developed the ATCC Genome Portal and ongoing whole-genome sequencing (WGS) initiative to address this problem; producing genomics data that can be traced back to the source material. Since the source material is taken straight from ATCC's repository, it has been authenticated and subjected to minimal passaging, which can cause lab-induced mutations.

There are several approaches for bioinformatics pipelines to assemble and annotate WGS data, however the processing of dozens of such assemblies in an automated, end-to-end fashion is often not discussed in detail. Here we present our own methodology that uses a hybrid of Illumina and Oxford Nanopore (ONT) sequencing technologies, and compares results to earlier internal assemblies as well as publicly available versions labelled as ATCC genomes.

In brief, materials in the repository are sequenced with Illumina MiSeq or NextSeq along with ONT GridION instruments. The FASTQs are checked for quality and filtered to remove low quality reads and adapters for Illumina and a minimum read length for ONT. Kraken2 is used to classify each read to check for contamination and to bin the reads to the appropriate taxonomic group. Reads are down sampled and assembled via Unicycler. Resulting contigs are further checked for quality and coverage statistics are generated. The

NCBI Prokaryotic Genome Annotation Pipeline (PGAP, installed locally) is used to annotate the genomes. Post-processing QC involves evaluating assembly statistics, completeness, improper assembly artifacts (such as inversions and unexpected repeats) and BLAST for a verification of genome identification (as some bacterial species can be very closely related). This pipeline ensures that the genomic data accurately represents the material sent by ATCC with traceability all the way to the original deposited source.

---PAGEBREAK---

Control Number: 2022-A-178-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Analysis of Complex Metagenomes with MetScale Workflows

Author Block: **M. Scholz**[1], N. Keplinger[1], C. Grahlmann[1], C. Hulme-Lowe[1], M. Isbell[1], T. Treangen[2], K. Ternus[1];
[1]Signature Science, LLC, Austin, TX, [2]Rice University, Houston, TX.

Abstract Body: Metagenomics is a powerful tool that allows researchers to gain insights into the taxonomic and functional content of complex microbial communities without the need for culturing. As metagenomics has expanded, there has been a corresponding increase in available taxonomic classification tools and reference databases to evaluate such sequences. However, positive calls at the strain, species, or genus levels may vary significantly among taxonomic classification tools. Furthermore, inconsistencies among the reference databases used by these tools can lead to false positives and/or false negatives, resulting in database-linked biases. To characterize and address these gaps, we developed MetScale workflows to allow users to execute multiple open source metagenomic tools and reference databases at one time and to assist users in differentiating true positive from false positive signals. MetScale can run online or offline on an air-gapped system, and its workflows include tools for Illumina read filtering, assembly, taxonomic classification, and functional inference (https://github.com/signaturescience/metscale). The MetScale final report summarizes the results for each sample analyzed and describes the species identified by different taxonomic classification tools and reference databases.

---PAGEBREAK---

Control Number: 2022-A-180-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Bioinformatic and Metabolomic Tools for Analysis of Toxicity of Single Cell Protein

Author Block: **S. M. Techtmann**[1], P. Kokate[1], L. Putman[1], L. Schaerer[1], T. K. Meyer[1], J. M. Pearce[2];
[1]Michigan Technological University, Houghton, MI, [2]Western University, London, ON, CANADA.

Abstract Body: There is an urgent need to address the imminent problem of food shortages. Single cellprotein (SCP), a form of alternative protein that is produced from microbial cells, could be a potential solution. One of the hurdles for the widespread use of SCP involves the

limitations of methods for prototyping SCP candidates. Part of this prototyping involves the methods for rapid assessment of the safety and potential toxicity of SCP products. Here we have developed a pipeline for screening the metabolites and genomic content of bacterial strains and communities in order to identify candidate cultures for single cell production from plastic waste.

At the metabolomic front, our pipeline allows computational assessment of our metabolite data to identify putative toxic compounds and provide a framework for assessment of safety based on chemicals produced by the SCP. Our open-source computational pipeline for assessment of metabolite toxicity includes three steps. First, we perform mass spectrometry analysis with MZmine 2. Then we assign formulas with MFAssignR, and finally we filter the data with ToxAssign. ToxAssign matches the formulas output by formula assignment to potentially toxic compounds in a local table, then look up toxic data on the Open Food Tox Database for the matched compounds. This allows for identification of potentially toxic metabolites in our SCP.

At the genome level, we have also developed a bioinformatic pipeline for assessment of potential toxins, allergens, and antibiotic resistance genes in SCP to identify potential candidates organisms. Our pipeline uses NNTox for identification of putative toxins, the AllergensOnline database for identification of potential allergens, and the CARD database along with the Resistance Gene Identifier for identification of antibiotic resistance genes. Our current pipeline has been tested on a microbial community enriched from compost and grown on deconstructed polyethylene terephthalate and high-density polypropylene. We have also tested this pipeline on various genomes of isolated bacteria. Our metabolite data has indicated that some members of the plastic degrading community could be producing potential toxins. Our bioinformatic approach has begun to identify which members of mixed communities may be producing potentially allergenic proteins and antibiotic resistance genes. This combination of chemical and genomic methods will allow for rapid assessment of suitability of candidate strains for SCP as well as inform community engineering for production of safe and nutritious alternative proteins.

---PAGEBREAK---

Control Number:	2022-A-183-NGS
Topic 1:	Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems
Topic 2:	
Publishing Title:	Exploring the Role of Bacterial Motility in Evolutionary Mechanisms for Antimicrobial Resistance
Author Block:	**L. Stabryla**, I. Keenum, J. Dootz, S. Servetas, J. Kralj; NIST, Gaithersburg, MD.
Abstract Body:	Antimicrobial resistance (AMR) poses a serious threat to global public health, food security, and development, and is considered one of the greatest systemic challenges of our time. Microbes have acquired resistance to every class of antibiotics, which challenges our ability to treat disease. However, our understanding of evolutionary and molecular mechanisms for AMR, particularly in complex bacterial communities, is limited which hampers our understanding of how resistant bacteria evolve and spread in the environment in ways we can track and predict with genomic-based approaches. Recapitulating evolution can reveal previously unidentified resistance mechanisms and emergence of new forms of resistance and thus help identify new targets for antimicrobials and surveillance monitoring. Of

particular interest is exploring the non-traditional role that flagellar motility may play in AMR and its physiologic response under antibiotic selection. Motility is a well-recognized fitness trait, has roles in virulence and pathogenicity, and has resulted in differential AMR profiles (i.e., microbes that swarm or have increased motility exhibit increased resistance to antibiotics), although its implication in AMR is poorly understood and may be attributed to the experimental approaches being used that provide a limited view of the AMR landscape. This work combines microbial evolution and integrated omics approaches to retrospectively understand the origin of resistance determinants that are clinically relevant today and identify new pathways, particularly those associated with motility, affected by various antimicrobials. Five independent replicate lineages of an avian *E. coli* strain were repeatedly exposed to subinhibitory concentrations of ampicillin and cephalexin over ten days. Various levels of resistance evolved among the replicates, with up to 64 and 32-fold increases in the minimum inhibitory concentration when exposed to ampicillin and cephalexin, respectively. Further, the higher resistances led to cross-resistance to other cephalosporins and extended spectrum penicillins as well as increased swim zones on motility agar. A detailed genomic analysis using whole-population genome sequencing and the breseq computational pipeline is now underway to identify point mutations arising in the evolved populations, and using transcriptomics, to see whether those genomic changes lead to sizeable changes in expression related to motility. Establishing motility as a marker for AMR has the potential to revolutionize measurement of AMR by adding physiologically-based measurement capabilities to our existing repertoire of molecular and chemical-based methods for AMR detection.

---PAGEBREAK---

Control Number: 2022-A-184-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: SeqScreen-LR: functional and taxonomic characterization of long read metagenomic data

Author Block: A. Balaji[1], **M. N. Nute**[1], B. Hu[1], A. D. Kappell[2], G. D. Godbold[3], K. L. Ternus[2], T. J. Treangen[1]; [1]Rice University, Houston, TX, [2]Signature Science, Austin, TX, [3]Signature Science, Charlottesville, VA.

Abstract Body: Affordable and accessible long read sequencing has opened the door to incorporating long read data into a wide variety of metagenomic analysis tasks, from obtaining high quality genome assemblies and binning to identifying structural variants. Though long reads offer better resolution than short-reads, assigning accurate functional and taxonomic labels to sequences is challenging due to a higher intrinsic error rate (especially for Oxford Nanopore Technology, or ONT). This presents a particularly challenging problem with respect to detecting and identifying closely related pathogens of interest. Here, we build upon our earlier tool, SeqScreen, and adapt it to identify Functions of Sequences of Concern (FunSoCs) from long read data. We show that on simulated and synthetic metagenomic data, SeqScreen-LR can identify multiple Open Reading Frames across the length of raw ONT reads and use it to accurately assign taxonomic labels using just a protein database. The taxonomic assignment is carried out using a combination of a majority voting heuristic and greedy weighted min-set cover approach. Further, we optimized SeqScreen-LR to run efficiently in a memory constrained environment (less than 32GB RAM), allowing it to be

utilized in limited resource settings. Lastly, SeqScreen-LR is also capable of processing batches of sequencing runs directly from the ONT MinION sequencer, enabling streaming functional and taxonomic profiling of metagenomic samples in a field setting.Affordable and accessible long read sequencing has opened the door to incorporating long read data into a wide variety of metagenomic analysis tasks, from obtaining high quality genome assemblies and binning to identifying structural variants. Though long reads offer better resolution than short-reads, assigning accurate functional and taxonomic labels to sequences is challenging due to a higher intrinsic error rate (especially for Oxford Nanopore Technology, or ONT). This presents a particularly challenging problem with respect to detecting and identifying closely related pathogens of interest. Here, we build upon our earlier tool, SeqScreen, and adapt it to identify Functions of Sequences of Concern (FunSoCs) from long read data. We show that on simulated and synthetic metagenomic data, SeqScreen-LR can identify multiple Open Reading Frames (ORFs) across the length of raw ONT reads and use it to accurately assign taxonomic labels using just a protein database. The taxonomic assignment is carried out using a combination of a majority voting heuristic and greedy weighted min-set cover approach. Further, we optimized SeqScreen-LR to run efficiently in a memory constrained environment (less than 32GB of RAM), allowing it to be utilized in limited resource settings. Lastly, SeqScreen-LR is also capable of processing batches of sequencing runs directly from the ONT MinION sequencer, enabling streaming functional and taxonomic profiling of metagenomic samples in a field setting.

---PAGEBREAK---

Control Number: 2022-A-189-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: TheiaProk: Species-Agnostic Bacterial Genome Analysis Workflows that Overcome Barriers to Bioinformatics for Public Health Laboratories

Author Block: **M. R. Scribner**, K. G. Libuit, R. A. Petit III, S. M. Wright, F. J. Ambrosio, C. J. Kapsak, E. A. Smith, E. L. Doughty, J. Sevinsky;
Theiagen Genomics, Highlands Ranch, CO.

Abstract Body: Whole genome sequencing of bacterial isolates can reveal a wealth of information for public health laboratories, including species identification, prediction of antimicrobial resistance determinants, and genetic relatedness of samples within outbreaks. However, these analyses often require numerous bioinformatics tools and are frequently restricted to implementation via the command line interface. We have developed an open source workflow called TheiaProk that enables rapid and high throughput analysis of bacterial genomes without the need to write or execute code.
TheiaProk is written in Workflow Description Language (WDL) and runs on Terra, a bioinformatics web application that provides a graphical user interface to open-source bioinformatics pipelines. Like many workflows for bacterial sequence characterization, TheiaProk performs quality assessment of sequencing reads followed by de novo assembly. The workflow subsequently implements tools for species identification and genome annotation including identification of antimicrobial resistance determinants. Species-specific sub-workflows proceed automatically based on taxonomic assignments using previously developed tools for organisms relevant to public health.
TheiaProk eliminates several barriers to analysis of bacterial whole genome sequencing

data. First, the workflow is open source, empowering public health scientists to critically evaluate and customize the parameters of their analysis. TheiaProk is also designed for implementation using a graphical user interface through Terra, which eliminates the need to run these bioinformatics tools via the command line interface. As such, large genomic data files are presented as links to their location in a cloud storage bucket within searchable data tables, and samples may be easily selected for analysis in parallel.
In addition, the TheiaProk workflow is designed to reduce the number of steps required between whole genome sequencing and obtaining valuable genomic insights. For example, the generation of a genome assembly through a de novo approach eliminates the need to identify a closely related reference genome and enables the user to run a single workflow for all bacterial specimens. Automatic initiation of species-specific workflows also saves the user from having to individually start these processes.
Stable releases of TheiaProk are currently used by numerous public health laboratories across the United States as we continuously expand and refine its analytical capabilities in order to reduce barriers to entry for bacterial whole genome sequencing analysis applications in public health.

---PAGEBREAK---

Control Number: 2022-A-191-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Performance of adaptor trimming algorithms and their effect on analysis of viral NGS data generated using Illumina iSeq and MiSeq platforms

Author Block: **G. Nabakooza**, D. D. Wagner, N. Momin, R. Marine, W. Weldon;
CDC, Atlanta, GA.

Abstract Body:
Performance of different adaptor trimming algorithms and their effect iSeq and MiSeq data and downstream analysis.Grace Nabakooza, Darlene D. Wagner, Nehalraza Momin, Rachel Marine, and William Weldon
Background: Next-generation sequencing (NGS) avails large genomic data for rapid disease surveillance. NGS-generated sequences often contain short fragments e.g., PCR primers and adaptors used during amplification and initial sequencing steps. Such contaminants reduce the data quality affecting downstream inferences. *Methods:* We surveyed studies of existing sequence trimmers selecting trimmers based on their sensitivity, specificity, positive and negative predictive values, and speed. The selected trimmers implement different algorithms: sequence matching (Trimmomatic and AdaptorRemoval v2), sequence overlapping (FastP), kmer-based (BBDUK), and probabilistic (SeqPurge). Then we sequenced the same SARS-Cov-2 (SC2) (n=8) and Noroviruses (n=7) samples using Illumina Iseq and Miseq platforms. The generated sequences per virus were adaptor and quality trimmed using the selected trimmers, and trimmer performance assessed based on the adaptor content and number, bases, quality score, and length of trimmed reads. Data were visualized and pairwise comparisons between trimmers were done using ggplot2 and Wilcoxon rank sum test in R v4.0.4, respectively.
Results: All trimmers returned clean paired reads with no adaptor contamination (<0.1%, MultiQC report). Except for Trimmomatic, Adaptor Removal, FastP, BBDUK, and SeqPurge had adaptors (<5.56%) in their trimmed single reads. Preliminary analysis of SC2 Miseq paired and single reads show significant differences in the mean number of reads, bases,

mean length, and reads (%) with >Q30 between raw and trimmed reads ($p < 0.00014$). However, BBDUK had a significantly higher percentage of reads with greater than Q30 than other trimmers (Wilcoxon rank sum test, $p < 3.7e^{-7}$). Paired reads showed significant differences in reads metrics between raw and trimmed reads (as above). Notably, BBDUK yielded significantly low number of paired reads relative to other trimmers ($p = 4.5e^{-6}$). Also, there were significant differences in the number of retained bases across all trimmers ($p < 1.7e^{-8}$), mean read length between Adaptor Removal and SeqPurge ($p = 0.017$), and the percentage of reads with >Q30 was highest in BBDUK trimmed paired reads ($p < 3.3e^{-6}$). *Conclusion:* Our preliminary data highlight the importance of adaptor and quality trimming for NGS quality. Optimal improvements in NGS quality metrics in turn facilitate downstream analyses such as genome assembly. Since many data cleaning tools are publicly available, performance of quality metrics and intended downstream analysis may serve as criteria for tool selection.

---PAGEBREAK---

Control Number: 2022-A-193-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: An Interactive Metagenomics Analysis Platform with Increased Accuracy and Precision at the Strain Level

Author Block: M. Narvaez, D. M. Walsh, K. Moffat, M. Dadlani, **N. Khan**;
CosmosID, Germantown, MD.

Abstract Body: Metagenomic sequencing is revolutionizing microbiology by facilitating rapid strain detection and discovery in a culture independent and unbiased manner. Accurate microbial identification from metagenomics sequencing is crucial for accurate downstream interpretation. Various methods for classification of metagenomic data, commonly referred to as metagenomics taxonomic classifiers, have been developed. Here, we perform a benchmarking study to evaluate the performance of the CosmosID HUB as compared to five other publicly available taxonomic classification algorithms: Centrifuge, Metaphlan3, Kraken2 Bracken, mOTUs2 and Metalign. These publicly available taxonomic classification algorithms are known for their high accuracy and precision when compared to other publicly available methods based on previous benchmarking evaluations. An ideal metagenomics classifier will properly identify a large number of microorganisms while displaying a small number of false positives at all taxonomic levels. For this evaluation, we used publicly available benchmarking datasets from CAMI2 (Mouse Gut Dataset; Sczyrba, 2017) and the *McIntyre et al* 2017 benchmarking paper. The CAMI2 dataset was designed for use as a common benchmarking tool in order to evaluate metagenomics pipelines in a standardized way and the McIntyre datasets consist of mock communities of known compositions. Overall, at species and strain levels, CosmosID HUB performs better than Centrifuge, Kraken2Bracken, Metaphlan3, Metalign and mOTUs2 across all evaluation metrics and particularly on the combined F1 score (the harmonic mean of sensitivity and precision). Except for Kraken2 Bracken, all the other remaining tools are unable to identify taxa to the strain level. When comparing Kraken2 Bracken to CosmosID HUB at strain level resolution, CosmosID HUB results clearly outperform Kraken2 Bracken and have better strain level identification. Most taxonomic profilers are unable to accurately identify bacteria to strain level because of the short reads mapping to multiple genomes due to either local or global

homology within the same species and different species as well. Unlike these pipelines, the CosmosID HUB's unique ability to differentiate between core and shared biomarkers among different prokaryotic genomes allows it to discriminate among strains of the same species accurately and precisely.

---PAGEBREAK---

Control Number: 2022-A-199-NGS

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Topic 2:

Publishing Title: Assessment of Differential Abundance Tools for NGS Microbiome Data: Towards Reproducibility and Robustness

Author Block: B. Ozdinc, K. Arogyaswamy, M. Dadlani, **D. M. Walsh**;
CosmosID, Germantown, MD.

Abstract Body: NGS methods have made possible accurate identification of a range of microbial taxa in a variety of sample types. However, the data generated from these methods are often large and require expertise in bioinformatic and statistical tools to determine which taxa are significantly associated with specific diseases, environments, or other variables. Several tools try to meet this need, but since each is designed to address specific sets of statistical challenges, they often provide conflicting results when applied to a single set of data. This divergence indicates that drawing accurate biological conclusions for any research question requires an understanding of which method is best suited to the structure of the data and the types of statistical challenges that may be present. To improve insight on tool-selection, we compared several differential abundance (DA) methods on synthetic datasets, mock communities, and publicly available whole genome shotgun data. Our synthetic datasets were designed *in silico* with specific differences in mean and standard deviation, creating standard comparisons with expected patterns of significant differentially abundant taxa to assess whether each tool performed as expected. Next, we tested the tools on *in vitro* mock communities with known relative abundances. These communities contain more of the variability introduced during sequencing and bioinformatic quality control, providing a bridge between cleaner *in silico* data and practical data with no known "ground truth." Finally, we tested our methods on real data sets. We compared results from ALDEx2, DESeq2, LEfSe, ANCOM BC, MicrobiomeDDA, and the Wilcoxon test. Our synthetic datasets included a set with abundances drawn from a single distribution to eliminate other tests that routinely found false positives. All six methods produced no differentially abundant taxa with this mock dataset, demonstrating robustness in avoiding spurious results. To determine which tools are most appropriate for each dataset, we plotted p value histograms of the output from each tool and determined the modal distribution among them. We used the consensus of those tools to generate a set of taxa likely to be biologically salient and meriting further investigation. We propose assessment of p value histograms as a method to assess the appropriateness of DA tools for specific datasets and, where possible, to use a consensus of taxa identified by multiple suitable tools to indicate statistically robust results.

---PAGEBREAK---

Control Number: 2022-A-200-NGS

Topic 1: Secret Ingredient: NGS to Uncover the Role of Microbes in Agricultural and Food Systems

Topic 2:

Publishing Title: Genome Sequencing and Molecular Characterization of a Novel *Canine Distemper Virus* Strain, Isolated from a Fox (*Otocyon megalotis*) in the United States

Author Block: **A. Roozitalab**, S. Kania, O. Elsakhawy, R. Donnell, M. Abouelkhair;
The University of Tennessee, KNOXVILLE, TN.

Abstract Body: *Canine morbillivirus* (*canine distemper virus*; CDV) is an RNA virus in the *Morbillivirus* genus of the *Paramyxoviridae* family. CDV is a highly contagious, systemic, and fatal disease that affects dogs all over the world, despite extensive and widespread vaccination in developed countries such as the United States. CDV endangers a wide range of wild animal populations and can cross species barriers hence represents a significant challenge at the wildlife-domestic animals interface. Understanding the evolution of emerging strains and the transmission risk from wildlife to domestic animals is highly important to mitigate the effect of spillover events on household animals. In this study, we present the genomic and phylogenetic analysis of a complete genome of a CDV strain from a one-year-old female bat-eared fox (*Otocyon megalotis*) in Tennessee, United States. RNA was isolated and cDNA synthesis was performed using NEB first and second strand synthesis kits. A library was prepared from cDNA using a Nextera™ DNA flex library preparation kit and sequencing was performed with an Illumina MiniSeq instrument, yielding greater than 980,000 150-base paired-end reads. SPAdes v3.14.0 was used to assemble the Illumina Miniseq short reads into an assembly graph using a variety of *k -mer* sizes. Phylogenetic analysis of the complete genomic sequences and separately the hemagglutinin gene sequence revealed a unique lineage circulating in US wildlife. More CDV strains from wild animals will be sequenced to gain a better understanding of the CDV genomic evolution as well as the likelihood of transmission from wildlife to domestic animals.

---PAGEBREAK---