Characterization of Microbiota in Cerebrospinal Fluid (CSF) from Patients with CSF Shunt Infections Using Shotgun Sequencing

**Background:** Treatment of hydrocephalus consists of surgical placement of a CSF shunt. Approximately 10% of patients develop a shunt infection within 1 year of CSF shunt placement. Approximately 20% of patients with first infection develop reinfection. It is not known whether reinfections are caused by an organism previously present in the host or are independent infection events. Identification of microorganisms associated with CSF shunt infections has traditionally relied on culture methods, but high throughput sequencing of 16S ribosomal RNA has been adopted more recently to identify bacterial species present. Here we present the results of a pilot study using whole genome shotgun sequencing and evaluate the additional resolution this method provides to our understanding of CSF shunt infection.

**Methods:** CSF samples were obtained from 6 patients having 2 infections, with one sample collected near the beginning and another near the end of each infection. The V4 region of 16S ribosomal RNA was amplified and sequenced. Alternatively, DNA was processed in duplicate by whole genome amplification (WGA) followed by shotgun sequencing. Taxonomic assignments of sequences obtained by 16S and WGA were compared against each other and with microbiological culture results. Non-human sequences from WGA were assembled and compared against known genomes from similar species.

**Results:** Taxonomic classification of bacteria observed by 16S and WGA was consistent with that obtained in the CSF cultures at the beginning of each infection episode. However, taxa assigned by 16S stopped at the genus level, and in one case (*Klebsiella pneumoniae*) 16S only identified the family (*Enterobacteriaceae*). WGA was able to identify all species detected in culture. Furthermore, WGA provided additional insights into the composition of the samples, such as showing that human DNA constituted 76 to 99% of the reads, identifying outlier samples of questionable quality, and detecting 2 cases of significant viral load. A few CSF samples produced a sufficiently large number of bacterial reads to allow partial assembly of the predominant species and comparison to known genomes to identify the closest matching strain.

**Conclusion:** This proof of concept study showed the value of shotgun sequencing in studying the microbiota of CSF shunt infections. Not only were the results consistent with culture-based methods, but additional insights could be gained regarding strain identity of predominant bacteria and identification of viral loads. This approach opens the door to a detailed understanding of the progression of infections and reinfections.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 86  
**Title:** GenEpiO and FoodOn: Enabling Data Interoperability for Infectious Disease Surveillance, Investigation and Control  
**Author Block:** E. Griffiths¹, D. Dooley², G. Gosal², N. Alikhan³, M. Sanchez⁴, T. Matthews⁵, A. Pertkau⁵, J. Adam⁵, R. Timme⁴, M. Graham⁵, G. Van Domselaar⁵, F. Brinkman¹, W. Hsiao⁶; ¹Simon Fraser University, Vancouver, BC, CANADA, ²University of British Columbia, Vancouver, BC, CANADA, ³University of Warwick, Coventry, UNITED KINGDOM, ⁴US Food and Drug Administration, College Park, MD, ⁵Public Health Agency of Canada, Winnipeg, MB, CANADA, ⁶BC Centre for Disease Control Public Health Laboratory, Vancouver, BC, CANADA.  
**Background:** The ability to share data between organizations is crucial for global, real-time infectious disease surveillance and investigation. Reliable capture and harmonization of whole genome sequencing (WGS) contextual information (sample source, experimental and bioinformatics methods, lab, clinical and epidemiological data) is critical for the interpretation of WGS results used for decision making in health crises. This data is often recorded using free text and institution-specific data dictionaries, requiring time-consuming and error-prone transformation before it can be used in investigations. Ontologies provide hierarchies of well-defined, standardized vocabulary enabling comparisons at different levels of granularity; universal IDs for disambiguating terms; built-in logic enhancing querying power; and synonyms that enable institutions to use preferred terms while linking to a standard, improving interoperability. We have created two ontologies to better harmonize and integrate genomics data into food microbiology and public health workflows, called the Genomic Epidemiology Ontology (GenEpiO) and the Food Ontology (FoodOn). Here, we describe the development of tools which facilitate ontology implementation within food safety and public health communities.  
**Methods:** User engagement activities identified vocabulary gaps, user needs, and use cases. Two ontology-driven tools were created to enable mapping of food microbiology and public health data to standardized terms. LexMapr, a Python-based, hybrid lexicon and rule-based system, was developed to address the many challenges in processing short textual data. Test datasets of metadata were mapped to GenEpiO and FoodOn to establish rules for natural language processing. Also, a Linux-based, open source, Python-driven web portal called the Genomic Epidemiology Entity Mart (GEEM) was developed to better enable the exchange of ontology-driven data specifications between agencies.  
**Results:** These tools and resources are currently being tested and evaluated for use in key databases and platforms for typing and tracking foodborne pathogens - Enterobase, GenomeTrakr and IRIDA. LexMapr testing indicates that the software has a high level of sensitivity in data clean-up, text matching and concept mapping. Furthermore, data specifications were created using GEEM for different applications, including an International Organization for Standards (ISO) standard for the implementation of WGS for food microbiology. The International GenEpiO Consortium (>80 members, from 15 countries) was also established to create consensus and uptake.  
**Conclusions:** The improved inferencing and computability of harmonized data provided by our resources and tools can enhance communication and analyses, resulting in faster hypothesis generation during investigations, and ultimately, better health outcomes.
The APHL and CDC Influenza Division began using the APHL Informatics Messaging System (AIMS) for transmission and analysis of next generation sequencing (NGS) data generated by the National Influenza Surveillance References Centers on behalf of CDC. With the success of this endeavor, laboratories have sought to leverage the data transmission and analysis capabilities for additional pathogens beyond influenza. This expansion has created the need for a “data parser” that can manage and route files based on data submitter, unique pathogen, and specific project to the appropriate data bucket and pipeline for analysis. The data parser, for its ability to slice AMD files and runs, is called Ninja. The Ninja software organizes and directs incoming NGS data based on pathogens, submitters, and projects to different services, e.g., placing the files into directories, rerouting files to another site, or notifying other processes about the availability of NGS data. The Ninja is able to sort and direct NGS data from multiple pathogens and projects into the appropriate data bucket and data analysis pipeline, based on the information users put on the sample sheet that will cue the Ninja to parse the NGS data into the appropriate places. Public health laboratories have a variety of hurdles in their use of NGS for disease detection and surveillance. Limited workforce and bioinformatics capacity are two of the biggest challenges. As public health laboratories expand their use of NGS, tools like Ninja will be instrumental in allowing them to use this technology for more Ninja’s technical capabilities allow public health labs to increase their efficiency and expand their use of NGS into a variety of pathogens.
Prediction of Antibiotic Minimum Inhibitory Concentration from Bacterial Whole Genome Sequence Data in *Klebsiella pneumoniae*

**Introduction:** Antimicrobial resistance testing has been a mainstay of clinical microbiology since the early 1970s. Phenotypic determination of minimum inhibitory concentration (MIC) is culture-dependent, requiring hours of growth before rendering an actionable result. Multiple studies have shown that decreasing the time between initial sample collection to actionable clinically relevant susceptibility results has multiple patient benefits, including decreased length of stay, decreased mortality, and decreased costs. Whole genome sequencing (WGS) has continued to decrease in cost while delivering faster results, proving useful for molecular microbiology. Recent advances in machine learning can develop classifiers that use bacterial WGS data to predict MIC within one dilution for many antibiotics. **Methods:** We used the whole genome sequence of 1,668 *K. pneumoniae* isolated from patients which had phenotypic antimicrobial susceptibility testing performed by BD Phoenix. We used these data to build classifiers using an XGBoost-based machine learning model to predict minimum inhibitory concentrations (MICs) for 20 antibiotics. These predictions were validated against a test set of isolates not included in the training set. **Results:** The overall accuracy of the model, within ±1 two-fold dilution factor, is 92%. Individual accuracies are ≥90% for 15/20 antibiotics tested. We show that the MICs predicted by the model correlate with known antimicrobial resistance genes. **Conclusion:** Importantly, the genome-wide approach described offers a method to predict MICs without knowledge of the underlying gene content. This study shows that machine learning can be used to build a complete *in silico* MIC prediction panel for *K. pneumoniae* and provides a framework for building MIC prediction models for other pathogenic bacteria. The ability to rapidly sequence bacterial genomes and then predict an MIC and resulting phenotype hours before culture-based methods have completed is a great potential advance for patient care and guiding empiric therapy.
**Introduction:** Infections caused by carbapenem-resistant Enterobacteriaceae (CRE) are associated with high mortality due to broad-spectrum antibiotic resistance. The plasmid-encoded *Klebsiella pneumoniae* carbapenemase (KPC) is the dominant mechanism of carbapenem resistance in the US. Both clonal expansion and horizontal transfer have been implicated in the spread of CRE. However, challenges sequencing plasmids have limited the ability to assign *bla*KPC to specific plasmid backbones to assess plasmid-mediated *bla*KPC transmission. Focusing on broad-host range IncN plasmids, which we previously detected in multiple strains of *Enterobacter cloacae* complex, we used MinION long-range sequencing to characterize and compare *bla*KPC-harboring plasmids in CRE clinical isolates collected at a tertiary care center where CRE are endemic. **Methods:** CRE isolates collected between 2010-2017 were identified on the basis of phenotypic resistance to meropenem (MIC≥2 mcg/dL) and sequenced using Illumina (n=469). *bla*KPC subtypes, multilocus sequence types, and plasmid replicon types were detected by SRST2 using the ARG-ANNOT, PubMLST, and PlasmidFinder databases, respectively. A subset of isolates found to have *bla*KPC-3 and a plasmid profile including an IncN replicon (n=15; 11 *K. pneumoniae*, 4 *E. cloacae*) underwent plasmid DNA extraction (Qiagen) followed by long-range sequencing using the MinION (Oxford). Hybrid plasmid assemblies were generated using SPAdes and visually curated and compared using Geneious (Biomatters). **Results:** We successfully localized both a plasmid replicon gene and *bla*KPC to a single contig for 8/15 isolates with median Illumina housekeeping read depths of 27.7 and 4,203 curated MinION sequencing reads (IQR 26.7-134.2 and 3,512.5-7,013.5, respectively). In 2 additional isolates, 2-3 large contigs mapped closely to a local internal reference plasmid (pNR0276, NCBI accession number PNXT00000000). *bla*KPC-3 was found on IncN plasmids in 6 isolates, including 3 *K. pneumoniae* from 3 different STs and 3 *E. cloacae* from 2 STs, ranging in length from 48,506-76,249 kb. In 4 *K. pneumoniae* isolates, *bla*KPC-3 was found on IncFI plasmids. Alignment of IncN plasmid sequences to pNR0276 indicated that 5/6 plasmids shared at least 90% pairwise identity over the full length of pNR0276, while one isolate harbored a truncated plasmid sharing an ~40 kb core region with pNR0276. **Conclusions:** Long-range sequencing enabled identification of an established *bla*KPC-3-harboring IncN plasmid backbone in carbapenem-resistant *K. pneumoniae* and *E. cloacae* at our hospital. Further study is needed to determine the extent of dissemination of IncN and other *bla*KPC-harboring plasmids among Enterobacteriaceae. Long-range sequencing has the potential to greatly facilitate comprehensive plasmid sequencing and demonstrate the important contribution of plasmids to the dissemination of CRE.
Abstract: We present Pan-synteny graphs, a multiple whole genome alignment model for understanding genome rearrangements and a graphical interface for browsing and understanding complex genome relationships. The visualization and interaction techniques demonstrate a powerful new kind of genome browser capable of summarizing information between hundreds of genomes. This effort touches on several different research fronts—graph representation of genomes and their alignments, synteny block analysis, whole genome sequence alignment, pan-genome analysis, multiple sequence alignment, and genome rearrangement analysis. Pan-synteny graphs represent a fundamentally new strategy to compare thousands of bacterial genomes in a scalable manner. Graph creation also identifies relative evolutionary events such as inversion, translocation, deletion, and insertion. Though this approach was originally developed from a pan-genome perspective for prokaryotes we are excited about its applicability to a wide range of topics. Algorithmically novel elements include the contextualization of synteny analysis both between and within multi-contig genomes. We also believe the algorithmic approach for discovering collision points has great value in the recognition of evolutionary relationships between a group of genomes. Pan-synteny graphs harness the information in pre-existing family databases, e.g. COGs and others. We will demonstrate how this information is able to make model construction more resilient to distant and complex evolutionary relationships as compared to existing tools such as Mauve and Harvest. This comparative graphical model also serves as a framework to analyze incomplete genomes. We hope to show that the graph abstraction and layout algorithm not only serve to make the resulting model approachable in terms of human cognition but represents a step forward in interactive comparative genomics.
Whose Lab is it, Anyway? --- Teaching Lab-specific Biases to a Metagenomics Taxonomy Classifier

J. A. Russell\(^1\), A. Shteyman\(^1\), D. Yarmosh\(^1\), P. Davis\(^1\), P. Li\(^2\), K. Davenport\(^2\), P. Chain\(^2\), J. Bagnoli\(^1\);
\(^1\)MRIGlobal, Gaithersburg, MD, \(^2\)Los Alamos National Laboratory, Los Alamos, NM.

Metagenomics is emerging as an important tool in biosurveillance, public health, and clinical applications. However, ease-of-use for execution and data analysis remains a primary barrier-of-entry to the full adoption of metagenomics in applied health and forensics settings. Here, we present PanGIA (Pan-Genomics for Infectious Agents), a novel framework for hosting, processing, analyzing, and reporting read-mapping data from metagenomics samples that can be run on commodity computer hardware. PanGIA was developed to address existing gaps that may preclude clinicians, medical technicians, forensics personnel, or other non-expert end-users from routinely leveraging metagenomics data for their needs. PanGIA is primarily meant for the detection and discovery of pathogenic microorganisms from clinical and environmental metagenomics data. PanGIA provides two forms of confidence scoring; the first pairs coverage data with ‘uniqueness’ information derived from each reference genome for a stand-alone determination of confidence for each query sequence at each taxonomy level, and the second compares a known ‘negative control’ profile with the profile of an unknown sample to determine significance in presence ‘above background’. Data can be quickly summarized within the graphical user interface to rapidly detect specific organisms-of-interest. PanGIA’s default parameters were optimized using a ROC-approach (Receiver Operating Characteristic curve) from in-silico-generated microbial communities. Recent work, leveraging a machine-learning approach, has explored the capacity of PanGIA to learn what known false-positives look like (across confidence score, normalized read abundance, reference genome linear coverage, depth-of-coverage, RPKM, and other metrics) such that PanGIA can more accurately distinguish potential false-positives in real-world laboratory sequencing data. In this way, over time and with initial user input, PanGIA can ‘learn’, recognize, and account for the contaminants and biases inherent to whichever laboratory it is placed in. This feature adds a unique level of confidence in discerning unambiguous detection events from low-confidence hits and false positives.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 92
Title: Development and Application Of QuAISAR-H: A Bioinformatics Pipeline for Short Read Sequences of Healthcare-Associated Pathogens
Author Block: R. A. Stanton, N. Vlachos, T. J. de Man, A. Lawsin, A. Laufer Halpin; Centers for Disease Control and Prevention, Decatur, GA.

Abstract: The application of whole genome sequencing (WGS) to surveillance projects and outbreak investigations of pathogens causing healthcare-associated infections (HAI) grants public health microbiologists an unprecedented level of resolution towards understanding the epidemiology of antimicrobial resistance, and transmission dynamics. However, the technical expertise required for processing and analyzing WGS data is often a major obstacle in public health laboratories, limiting the feasibility of implementing WGS on a wide-scale. Furthermore, healthcare-associated pathogens are uniquely challenging because of their diversity; our group alone has sequenced more than 50 different species causing HAIs. Finally, the lack of established standards for performing sequence analysis has left a gap in public health practice. To address these shortcomings, we have developed QuAISAR-H: a specialized pipeline for Quality control, Assembly, species Identification, Sequence typing, Annotation, and Resistance mechanisms for Healthcare-associated pathogens. QuAISAR-H currently runs on the CDC’s high performance computing cluster, utilizing open source software and custom scripts. It accepts and is optimized for raw reads generated by Illumina short read sequencers and initially performs a variety of quality control assessments, including species identification and contamination checks using Kraken and Gottcha. Genome assemblies are generated using SPAdes, classified using MLST definitions and functionally annotated by Prokka. Antimicrobial resistance genes are identified using multiple databases from both the raw reads (using SRST2) and the assemblies (using c-SSTAR). The output assemblies and high-quality, cleaned reads generated by QuAISAR-H can be used for downstream phylogenetic analysis. The implementation of QuAISAR-H has allowed us to move towards a more standardized approach of analyzing WGS data from HAI pathogens. We have iterated and streamlined the pipeline through processing more than 3400 isolates sequenced internally and externally, including those from 45 HAI outbreaks. A graphical user interface to provide public health laboratories across the country with direct and easy access to QuAISAR-H is currently under development and will be available through the CDC’s Office of Advanced Molecular Detection online portal. This will enhance not only local capacity, but also national efficiency in utilizing WGS data for HAI surveillance and investigation.
Abstract: GenomeTrakr proficiency testing for foodborne pathogen surveillance

Pathogen monitoring is becoming much more robust as sequencing technologies become more affordable and accessible worldwide. This transition is especially apparent in the field of food safety, which has demonstrated how whole genome sequencing (WGS) can be used on a global scale to protect public health. GenomeTrakr coordinates the WGS performed by public health agencies and other partners by providing a public database with real-time cluster analysis for foodborne pathogen surveillance. As growing numbers of public health labs use WGS technology to support enforcement decisions, it is essential to have confidence in the quality of the data being used and the downstream data analyses which guide these decisions. Routine proficiency tests, such as the one described here, have an important role in ensuring the validity of both data and procedures. GenomeTrakr ran an annual internal proficiency test through 2015 that is now harmonized with PulseNet. In 2015 the GenomeTrakr proficiency test consisted of 8 isolates of common foodborne pathogens; participating laboratories were required to follow a protocol to culture these and perform WGS. Resulting sequence data were evaluated for proper annotation, sequence quality, and applicability to downstream bioinformatics analyses. Overall, this exercise revealed the degree of variation which should be expected in sequence data produced across a diverse network of laboratories. Illumina MiSeq sequence data collected for the same set of strains across 21 different labs exhibited high reproducibility, while revealing a narrow range of technical and biological variance. The numbers of SNPs reported for sequencing runs of the same isolates across multiple labs support the robustness of our cluster analysis pipeline in that each individual isolate cultured and resequenced multiple times in multiple places are all easily identifiable as originating from the same source. Subsequent proficiency tests confirm these results.
Impact of Antibiotic and Innate Immune Pressures on Enterococcal Adaptation in the Human Bloodstream

D. Van Tyne¹, A. L. Manson², M. M. Huycke³, J. Karanicolas⁴, A. M. Earl², M. S. Gilmore⁵;
¹University of Pittsburgh School of Medicine, Pittsburgh, PA, ²Broad Institute, Cambridge, MA, ³University of Oklahoma Health Sciences Center, Oklahoma City, OK, ⁴Fox Chase Cancer Center, Philadelphia, PA, ⁵Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA.

Multidrug-resistant enterococci emerged in the early 1980s, and are now among leading causes of drug-resistant bacterial infection worldwide. We used functional genomics to study one of the earliest outbreaks of multidrug-resistant Enterococcus faecalis bacteremia, to determine how a clonal lineage adapted to grow and survive in the human bloodstream. Genome sequence analysis of 62 closely related strains revealed a progression of increasingly fixed mutations, as well as repeated independent occurrence of mutations in a relatively small set of genes. The most frequently encountered independent mutations we observed occurred in a novel pathway that rendered E. faecalis better able to withstand antibiotic pressure and innate defenses in the bloodstream, and were associated with changes in cell surface-associated polysaccharides. A shift in mutation pattern then occurred, which corresponded to the introduction of carbapenem antibiotics in 1987. This work uncovers new pathways that allow enterococci to survive the transition from the gut into the bloodstream, positioning them to cause infections associated with high mortality.
Characterization of Tissue-associated Metagenomes Using Selective Nanopore Sequencing

J. Wang, C. Jones, T. Furey, S. Sheikh, O. Finkel, J. Dangl; University of North Carolina at Chapel Hill, Chapel Hill, NC.

While 16S rDNA profiling has been the standard approach to characterizing host-associated microbiome communities, it produces taxonomic classifications practically limited to the genus level and suffers from PCR and other biases. Whole metagenome sequencing produces more specific taxonomic information and an estimate of genetic content describing the functional capacity of a microbial community. However, metagenomic studies are expensive and require high sequencing depth, especially in tissue-associated microbiomes, where host DNA makes up the vast majority of the sequenced reads (90-99+%). We describe a real-time sequencing and analysis approach using Oxford Nanopore sequencers that enables real-time enrichment or depletion of specific sequences. Using the "read-until" functionality of the MinION sequencer, we perform basecalling and alignment of partial read sequences in real time on a distributed cloud computing platform and eject reads belonging to the host genome, thereby increasing the relative and absolute abundance of microbial sequences. This approach is essentially unbiased compared to existing method for preferential cell lysis and DNA extraction, and produces an actual increase in sequenced microbial DNA unlike post-sequencing filtering. We demonstrate the power of this approach by depleting host (human) DNA in a mock host-microbial metagenome, and in a colon biopsy sample to describe the composition and function of the mucosa-associated microbiome in the colon. We observed a two to six-fold increase in the relative abundance of microbial sequences relative to host, depending on the initial proportion. This selection method produces no detectable false-positive depletion (of microbial sequences) or selection bias in the retained reads. We additionally propose a simple and effective method for accurately classifying observed long reads as host, or to their appropriate species/strian-level taxa. These host-depleted metagenome experiments - with novel methods to efficiently classify long, error-prone reads - demonstrate the power of tightly coupled sequencing and informatics protocols to enable efficient investigation of disease-relevant tissue-associated microbiota.
Clinical Whole-Genome Sequencing of *mycobacterium tuberculosis* complex Isolates - 2½ Years of Experience Analyzing, Reporting and Improving TB Testing in New York State

**K. Musser**¹, J. Shea¹, P. Lapierre¹, T. Halse¹, J. Lemon², J. Rakeman³, V. Escuyer¹; ¹Wadsworth Center, NYSDOH, Albany, NY, ²Public Health Laboratory, New York City Department of Health and Mental Hygiene, New York City, NY.

**Background:** *Mycobacterium tuberculosis* (MTB) is an important pathogen, infecting more than a third of the world population; New York State (NYS) has the 3rd highest number of cases by state in the US. The cost and time associated with diagnostic testing and treatment of MTB can be considerable and weeks to months are required to identify, assess drug susceptibility, and generate molecular genotypes. Our laboratory developed and validated a comprehensive whole-genome sequencing (WGS) assay to characterize MTB complex (MTBC) isolates, replacing seven molecular tests. We implemented this testing in March of 2016 and have continually measured its performance, assessed turnaround time (TAT), success at resistance prediction and high-resolution genotyping.

**Methods:** The MTBC WGS assay is comprised of a novel DNA extraction, optimized library preparation, paired-end WGS, and an in-house developed bioinformatics pipeline; numerous quality control steps are incorporated in this testing. Following DNA sequencing, the pipeline performs analysis using three principal components: modules used for the phylogenetic analysis, modules used to perform taxonomic identification of the samples, and modules used for SNP calling and resistance profiling. The results from all three components generate a final comprehensive report for each sample analyzed that is reported through our LIMS.

**Results:** To date we have tested 1634 MTBC strains from unique NYS patients, including NYC. Of these, 5 members of MTBC have been identified: 1560 *M. tuberculosis*, 27 *M. bovis*-BCG, 26 *M. bovis*, 17 *M. africanum*, and 4 *M. orygis*. In-silico spoligotypes were generated for 96.5% of strains tested, and strains found to be closely related (<20 SNPs genome wide) were reported for epidemiological investigation. Resistance profiles of the MTBC strains showed 79.8% to be susceptible to eight drugs, 7.8% resistant to at least isoniazid, 2.4% multidrug resistant (MDR), and 0.12% extensively drug resistant (XDR) strains. When compared with conventional phenotypic drug susceptibility testing (DST), our assay was found to have an overall resistance predictive value of 94% and a susceptibility predictive value of 98% based on >8000 phenotype-genotype comparisons. We have assessed TAT since implementation and reduced our initial 8-day TAT to 5 days from MTBC DNA extraction to report. This TAT has resulted in genotypic resistance predictions being reported an average of 8 days earlier than first-line phenotypic DST.

**Conclusions:** This TB WGS clinical assay is providing comprehensive detection of drug resistance, identification to the MTBC member and typing for epidemiological investigations. As a result of improvements as well as updates to analyze more samples at one time, an improved TB WGS pipeline is in use. This assay continues to improve patient management and is supporting epidemiological investigations in NYS and NYC.
Title: NGS Applied to the Epidemiology of Influenza a Virus Diversity in Brazil

A. B. Veiga¹, T. Song², T. G. Baccin¹, T. S. Gregianini³, H. V. Bakel⁴, A. García-Sastre⁵, E. Ghedin²;

¹Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, BRAZIL, ²New York University, New York City, NY, ³Laboratório Central de Saúde Pública da Secretaria de Saúde do Estado do Rio Grande do Sul – LACEN/SES-RS, Porto Alegre, BRAZIL, ⁴Mount Sinai Hospital, New York City, NY, ⁵Icahn School of Medicine at Mount Sinai, New York City, NY.

Abstract: Systematic surveillance of seasonal influenza A viruses using next generation sequencing (NGS) has the potential to contribute to early detection of novel influenza strains in the human population. In this study, we sequenced and analyzed clinical samples collected in Brazil between 2009 and 2016 from 220 individuals infected with pandemic H1N1 (H1N1pdm09) or H3N2 influenza A virus. Phylogenetic analyses show persistence of strains from one season to the next in Brazil, with introductions of new strains from global circulating viruses. An analysis of single nucleotide variants (SNV) in the NGS data reveals mixed infections with minor circulating strains that also appear to seed the next season. Some SNVs are located in antigenic sites of the hemagglutinin, leading to changes in antigenicity in recent strains. For example, the non-synonymous mutation A538C in segment 4 of H1N1pdm09 (K180Q substitution in the HA antigenic site) appeared in strains during the 2013-2014 influenza season in the Northern Hemisphere, but the SNV analysis shows that minor variants carrying this mutation had been circulating as early as 2011. In 10 of the 220 infected individuals sampled we also detected mixed subtype infections, considered a rare occurrence in the human population. NGS combined with minor variant analysis proves to be a powerful surveillance tool to identify mixed infections and potential circulating strains in upcoming seasons.
**Background:** Food induced infectious diseases still remain a major cause of health problems across the globe. With the continuously increased use of next-generation sequencing (NGS) in the field of infectious disease outbreak analysis and food safety controls, there is a strong need for fast turnaround time from sample arrival to analysis results. While runtime of data analysis software has significantly decreased, the overall turnaround time from sample arrival to interpretable analysis results remained nearly the same due to the sequential paradigm of data production and analysis. To overcome this limitation, we developed a collection of tools for sequence analysis while the sequencer is still running.  

**Methods:** The presented methods include software for read mapping (HiLive; Lindner et al., 2017, doi:10.1093/bioinformatics/btw659), taxonomic classification (LiveKraken; Tausch et al., 2018, doi:10.1093/bioinformatics/bty433), privacy preservation (PriLive; Loka et al., 2018, doi:10.1093/bioinformatics/bty128), pathogen identification (PathoLive; Tausch et al.) and a workflow for SNP/variant calling (Loka et al.) while the sequencer is running.  

**Results:** We are able to show that each of our tools generates comparable or superior results to established tools in the named fields. HiLive’s accuracy (F1 = 0.761) is slightly higher than that of the other tested approaches (BWA: 0.760, Bowtie2: 0.742) with the end of a sequencing run. LiveKraken performs identical to Kraken with the end of a full MiSeq run, while reaching comparable accuracy after less than half of a run (F1 = 0.96 at cycle 80 of 216). PriLive filters human reads more accurately (F1 = 99.961) than BMTagger (99.956) and DeconSeq (99.941) and can moreover mask sensitive data before it is completely produced. PathoLive combines a live mapping approach with novel background masking techniques and thereby achieves highest accuracy on a real HiSeq run (ROC-auc = 0.97 after 36h turnaround time) compared to Clinical PathoScope (ROC-auc = 0.91 after 95h turnaround time) and Bracken (ROC-auc = 0.48 after 95h turnaround time).  

**Conclusion:** With each of these tools, we prove the ability to generate meaningful results with or even before the end of a sequencing run. This allows minimizing the turnaround time of a variety of tasks and can thereby increase the efficiency of high throughput routine analyses. It could furthermore significantly reduce the response time in urgent cases of infectious disease outbreaks. Since more and more institutions have their own sequencers available, the parallelization of wet- and drylab is at hand.
Beaver fever: Whole Genome Characterization of Waterborne Giardia Isolates Revealed Mix Assemblages and Zoonotic Transmission

K. Tsui¹, R. Miller², M. Uyaguari-Diaz², P. Tang¹, C. Chauve³, W. Hsiao², J. Isaac-Renton², N. Prystajecky²;

¹Sidra Medicine, Doha, QATAR, ²University of British Columbia, Vancouver, BC, CANADA, ³Simon Fraser University, Vancouver, BC, CANADA.

Giardia causes the diarrheal disease known as giardiasis; transmission through contaminated surface water is common. The protozoan parasite's genetic diversity has major implications for human health and epidemiology. To determine the extent of transmission from wildlife through surface water, we performed whole-genome sequencing (WGS) to characterize 89 Giardia duodenalis isolates from both outbreak and sporadic infections: 29 isolates from raw surface water, 38 from humans, and 22 from veterinary sources. Using single nucleotide variants (SNVs), combined with epidemiological data, relationships contributing to zoonotic transmission were described. Two assemblages, A and B, were identified in surface water, human, and veterinary isolates. Mixes of zoonotic assemblages A and B were seen in all the community waterborne outbreaks in British Columbia (BC), Canada, studied. Assemblage A was further subdivided into assemblages A1 and A2 based on the genetic variation observed.

Abstract:

The A1 assemblage was highly clonal; isolates of surface water, human, and veterinary origins from Canada, United States, and New Zealand clustered together with minor variation, consistent with this being a panglobal zoonotic lineage. In contrast, assemblage B isolates were variable and consisted of several clonal lineages relating to waterborne outbreaks and geographic locations. Most human infection isolates in waterborne outbreaks clustered with isolates from surface water and beavers implicated to be outbreak sources by public health. In-depth outbreak analysis demonstrated that beavers can act as amplification hosts for human infections and can act as sources of surface water contamination. It is also known that other wild and domesticated animals, as well as humans, can be sources of waterborne giardiasis. This study demonstrates the utility of WGS in furthering our understanding of Giardia transmission dynamics at the water-human-animal interface.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 100
Title: Refactoring the NCBI Prokaryotic Genome Annotation Pipeline into a Stand-alone Tool
Author Block: F. Thibaud-Nissen¹, D. Slotta¹, A. Badretdin¹, B. Kiryutin¹, A. Gourianov¹, B. Busby¹, R. Cohen¹, W. Hlavina¹, M. Hsieh², S. Turner²;
¹NCBI/NLM/NIH, Bethesda, MD, ²Pacific Biosciences, Menlo Park, CA.
Abstract: The NCBI Prokaryotic Genome Annotation Pipeline (PGAP) has been used to annotate RefSeq prokaryotic genomes since the early 2000s, increasing in quality and consistency over the years. PGAP annotation, also offered as a service to researchers submitting genome assemblies to GenBank has become a reliable resource for the prokaryotic community. We have re-factored PGAP into a stand-alone pipeline that can be executed outside of NCBI on individual computers or in a cloud environment. The pipeline is written in CWL, which executes programs wrapped in Docker containers, to run on a variety of platforms. To ensure conformance of the stand-alone results with results generated at NCBI, manually curated evidence and other datasets used by the pipeline are bundled and distributed with the pipeline. The goal is for stand-alone PGAP to produce annotation that is in line with internal NCBI PGAP and that is submittable to GenBank. We expect that making PGAP portable will accelerate research by providing scientists a quality annotation of the genomes they assemble prior to submission. It will also give users an opportunity to iterate over the assembly process until the assembly quality is high enough to produce quality annotation. We will describe the stand-alone PGAP prototype and the results of the annotation tests performed on multiple platforms and by multiple users across several locations, with respect to performance and conformance.
Diagnosis and Characterization of Canine Distemper Virus Through Real Time Sequencing by MinION Nanopore Technology

Istituto Zooprofilattico dell' Abruzzo e del Molise, Teramo, ITALY.

Rapid identification of the etiologic agent of an infectious disease is essential for setting up treatment and preventive measures. In general, pathogen identification is performed by direct diagnostic tests which normally include amplification of target nucleic acid by PCR-based assays. Although these approaches are highly specific and, often, validated, they suffer a number of limitations, including the difficulties of testing for the plethora of rare pathogens that might be expected to cause a given pathology and their inability to identify new or unexpected pathogens, eventually originated from cross-species jumps. Therefore, the existence of other more rapid, broad-range and sensitive techniques have become more and more important in the milieu of laboratory diagnosis of infectious diseases. In this perspective, nucleic acids purified from the brain tissue of a dog succumbed after severe neurological signs were processed with the MinION (Oxford Nanopore Technologies, Cambridge UK) sequencing technology. Canine distemper virus (CDV) infection was diagnosed. The earliest detection of sequence reads belonging to CDV was accomplished within the first 20 minutes of real time sequencing. Subsequently, a specific real time RT-PCR assay and immunohistochemistry were used to confirm the presence of CDV RNA and antigen, respectively, in tissues. This study supports the use of the MinION in veterinary clinical practice with tremendous advantages in terms of rapidity and accuracy of molecular diagnosis.
Background: Whole genome sequencing (WGS) is rapidly being adopted by Public Health in many jurisdictions, creating a need for rapid, robust analytical tools. Single nucleotide polymorphism (SNP) genotyping panels have been developed for numerous organisms based on canonical SNPs that are discriminatory for clonal populations. Using canonical SNP panels, new isolates can rapidly be placed within the population structure without the need to rebuild phylogenetic trees for the entire population. A canonical SNP-based nomenclature can facilitate long-term surveillance by allowing numerous comparisons of isolates across time in a context broader than typically considered for outbreak response.

Methods: Biohansel rapidly classifies WGS data into hierarchical subtypes without the need for assembly. Canonical SNP schemas for two prevalent Salmonella serovars (S. Enteritidis and S. Heidelberg) have been incorporated into biohansel, and user-defined schemas can also be supplied at runtime for subtyping other pathogens. Biohansel identifies SNPs using the Aho-Corasick algorithm (Ju et al., 2017, doi.org/10.1101/229708) according to defined k-mers containing target SNPs. Results are evaluated using a quality assurance module which identifies problematic samples according to the number of targets found, target coverage, and concordance with the population structure defined by each schema. Possible mixed samples are identified based on the presence of discordant sets of SNPs and presence of multiple SNPs for each target. Biohansel is a Python 3 application and available on PyPI, Conda and as a Galaxy tool. Source code is available at https://github.com/phac-nml/biohansel.

Results: We demonstrate the utility of biohansel by rapidly analyzing >23,000 S. Enteritidis and >3,000 S. Heidelberg WGS datasets from public repositories using minimal computational resources, and by identifying subtype associations with commodities and geography. Biohansel proved useful to rapidly identify closely related isolates and exclude poor quality WGS datasets, enabling the creation of reference-mapped phylogenetic trees with the high discriminatory power needed for traceback investigations. The tool was also able to detect and subtype Salmonella in shotgun metagenomics datasets obtained from clinical stool samples. The Aho-Corasick algorithm for k-mer searching is as fast as NCBI BLAST+ against assembly contigs (~0.4s) and is 10 times faster than Jellyfish (~33s vs ~356s) for typically sized Salmonella read sets (30-100X coverage).

Conclusions: In a public health context, biohansel enables rapid and high resolution classification of North American isolates, providing a robust, stable framework for source attribution and supporting identification of possible interventions to reduce contamination of food products.
Background: Whole genome sequencing (WGS) is a powerful tool for public health infectious disease investigations owing to its higher resolution, greater efficiency and cost-effectiveness over traditional genotyping methods. However, implementation of WGS in routine public health microbiology labs is impeded by the complexity in data management, availability of easy-to-use pipelines, integration of pipeline results with epidemiological metadata, and restrictive jurisdictional data sharing policies. To address these issues, we developed the Integrated Rapid Infectious Disease Analysis (IRIDA) platform—a user-friendly, decentralized, open source bioinformatics and analytical web platform—to support real-time infectious disease outbreak investigations using WGS data.

Methods/Results: IRIDA stores and manages WGS data alongside contextual metadata—providing a single system for processing and generating reports on sequenced samples. WGS data is automatically uploaded to IRIDA using a tool installable on a sequencing instrument. Data is then processed to evaluate quality, assemble, perform in silico sequence typing, and save results into the epidemiological metadata system. Typing of Salmonella genomes uses SISTR, a tool for Salmonella serovar prediction and cgMLST analysis from WGS data. Additional k-mer based typing pipelines include MentaLiST for cg/wgMLST and biohansel for SNP-based typing. SNVPhyl provides whole genome phylogenetic analysis using SNV/SNPs; Mash provides rapid distance estimation to existing genomes in RefSeq. Genomes may be sent to IslandViewer for genomic island detection. Pipelines may be configured to trigger automatically on upload of new WGS data, or users may select sets of samples for additional analysis through the IRIDA pipelines.

The IRIDA metadata system integrates data generated from a pipeline—such as sequence type—with user-provided metadata into a single table. Users may toggle the display of metadata fields and save specific views of the metadata for later use. These views of metadata may also be visualized alongside a phylogenetic tree. The IRIDA REST API enables secure exchange of genomic and epidemiological metadata, enabling construction of a decentralized genome data sharing network. The IRIDA REST API may also be used to extend IRIDA’s functionality, such as through additional tools for custom report generation or integration with the phylogeographic software GenGIS.

Conclusion: IRIDA is successfully deployed as the official bioinformatics platform for public health genomics in the pan-Canadian Public Health Laboratory Network (CPHLN). The storage, management, and analysis of WGS data alongside contextual metadata has helped simplify surveillance and outbreak investigation activities. IRIDA is open source and freely available at https://github.com/phac-nml/irida and http://irida.ca.
Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations

A. Pightling, J. Pettengill, Y. Luo, J. Baugher, H. Rand, E. Strain; U.S. Food and Drug Administration, College Park, MD.

Whole-genome sequence (WGS) analysis has revolutionized the food safety industry by enabling high-resolution typing of foodborne bacteria. Higher resolving power allows investigators to identify origins of contamination during illness outbreaks and regulatory activities quickly and accurately. Government agencies and industry stakeholders worldwide are now analyzing WGS data routinely. Although researchers have published many studies that assess the efficacy of WGS data analysis for source attribution, guidance for interpreting WGS analyses is lacking. Here, we provide the framework for interpreting WGS analyses used by the Food and Drug Administration’s Center for Food Safety and Applied Nutrition (CFSAN). We based this framework on the experiences of CFSAN investigators, collaborations and interactions with government and industry partners, and evaluation of the published literature. A fundamental question for investigators is whether two or more bacteria arose from the same source of contamination. Analysts often count the numbers of nucleotide differences (single-nucleotide polymorphisms [SNPs]) between two or more genome sequences to measure genetic distances. However, using SNP thresholds alone to assess whether bacteria originated from the same source can be misleading. Bacteria that are isolated from food, environmental, or clinical samples are representatives of bacterial populations. These populations are subject to evolutionary forces that can change genome sequences. Therefore, interpreting WGS analyses of foodborne bacteria requires a more sophisticated approach. We present a framework for interpreting WGS analyses that combines SNP counts with phylogenetic tree topologies and bootstrap support. We also elucidate the roles of WGS, epidemiological, traceback, and other evidence in forming the conclusions of investigations, making clear that WGS data alone is insufficient for links between bacterial isolates to be made. Finally, we present examples that illustrate the application of this framework to real-world situations.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 106
Title: GenomeTrakr Database and Network: WGS Network for Real-Time Characterization and Source Tracking of Foodborne Pathogens

Abstract:
A national database of federal, state, academic and international laboratories has been using WGS data to rapidly characterize pathogens. This mature GenomeTrakr network is part of NCBI Pathogen Detection website. Public health agencies (FDA, CDC and USDA-FSIS) collect and share data in real time. This high-resolution, rapidly growing database is actively being used in outbreak investigations at state, national, and international levels. 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 new “open data” model allows greater transparency between federal/state agencies, industry partners, academia, and international collaborators. This database has continued to grow and diversify the foodborne pathogen database doubling in the last year to ~207,000 draft genomes. Two new international surveillance efforts were added to collect food, animal and environmental isolates including Campylobacter. NCBI has released new data analysis tools that improve rapid interpretation and visualization. NCBI, currently is producing daily clustering results for 22 pathogens including: Salmonella, Listeria, E. coli, and Campylobacter. The high-resolution WGS signal in concert with epidemiological or inspection evidence has drastically enhanced our ability to identify the food sources of current outbreaks for foodborne pathogens with ~200 regulatory clusters examined in 2017. Results demonstrate global benefits of having an open data model. Understanding root causes of foodborne contamination assists our academic, public health and industry partners to develop preventative controls to make food safer globally.
Ultra-Rapid Sample-to-Answer for Fieldable Genomic Sequencing-Based Biothreat Detection

T. Reed¹, M. Karavis², S. Deshpande³, R. Lewandowski¹, C. Anderson², M. LaFrance¹, P. Roth⁴, A. Liem⁴, R. C. Bernhards⁵;

¹CBRNE Analytical & Remediation Activity, 20th CBRNE Command, US Army, Aberdeen Proving Ground, MD, ²Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD, ³Science & Technology Corp. support to Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD, ⁴DCS Corp. support to Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD, ⁵Defense Threat Reduction Agency, Ft. Belvoir, VA.

Rapid and accurate detection technologies are critically needed in the field, especially for unknown, emerging, and genetically modified biothreats. Next-generation sequencing (NGS) technologies are superior in that the entire genome can be analyzed, which allows for unbiased, conclusive identification, and the ability to detect new and synthetically modified threats. However, most NGS sequencing technologies have substantial size, power, and sample preparation requirements which severely limits their use in far-forward environments, and current methodologies for sample-to-answer take multiple days to complete. The MinION nanopore sequencer developed by Oxford Nanopore Technologies has recently emerged as a portable NGS technology. Nanopore sequencing utilizes biological proteins as nanopores for the passage and identification of DNA and RNA molecules. Improvements in error rates combined with the high amount of read generation have made nanopore sequencing comparable to existing NGS sequencing technologies, such as Illumina and PacBio, without the need for large, expensive equipment. The MinION is able to fit in the palm of your hand, offering the capability to conduct true field-deployable sequencing, which can allow for rapid identification of unknown threats and disease monitoring in resource-limited settings. This project aims to accelerate the time from sample to answer, simplify the procedures, and reduce equipment/power that is needed. An optimized workflow was established using simple, fieldable, and rapid sample and library preparation procedures. The workflow includes the use of the portable OmniLyse bead beading device capable of lysing spores within two minutes, a rapid DNA purification protocol, and an eight minute library preparation. In addition, the utility of the VoITRAX automated sample/library preparation device, the Flongle flow cell adapter, and the MiniTI miniature processor are currently being investigated for inclusion into the workflow. Within 10 minutes of sequencing on the MinION, enough reads are generated to conclusively identify the organisms present in the sample. Automatic offline live basecalling is used during the sequencing run, and after sequencing is complete, the data is analyzed instantly using offline software developed at ECBC. Using this workflow, raw sample-to-answer can be achieved in approximately one hour. Field demonstrations are being conducted with DoD mobile lab operators for assessment. The goal is to allow for genomic sequencing identification to be performed rapidly by minimally trained personnel in low-resource environments and without the need for high-powered lab equipment. Using this procedure, the MinION could be used by the warfighter to rapidly identify unknown biothreats on the battlefield or in expeditionary analytic scenarios.
Session: **Poster Session B**
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 109
Title: Tracing the Origins of Hospital-onset *Clostridioides difficile* Infections

**Author Block:**
- J. Worley¹, C. Cummins¹, M. Delaney¹, A. Dubois¹, S. Men¹, M. Klompas², L. Bry¹;

¹Massachusetts Host-Microbiome Center, Department of Pathology, Brigham and Women's Hospital, Boston, MA, ²Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA.

**Background:** *Clostridioides difficile* is a leading cause of health care-associated infection in the United States and the leading cause of death from a gastrointestinal pathogen. The annual costs associated with its treatment are frequently estimated to be in excess of $5 billion. We designed an experiment to address if patients who develop hospital-onset *C. difficile* infection (CDI) were infected by strains commonly found at the hospital, from other patients, or asymptptomatically carried upon admission. While some strains have been more common, particularly sequence type 1/NAP1/PCR ribotype 027 (ST1), there is high genetic diversity within disease-causing *C. difficile*. Whole-genome sequencing, which can identify clonally related bacteria, was used to address this question by sequencing strains from CDI presenting patients and an incoming patient screen.

**Methods:** The study period was September 2017 through May 2018. Patients admitted to the intensive care units are screened for vancomycin-resistant Enterococci by rectal swab (VRE swab) upon admission and weekly thereafter. These swabs were screened for *C. difficile* to identify strains arriving to the hospital. VRE swabs were collected from November through April 15th, while stool was collected over the entire period. Stool collection was hospital wide. Isolates were sequenced using the Illumina MiSeq platform. Single-nucleotide polymorphisms were identified de novo using kSNP and through core-gene alignment. Sequence types and genetic features were assessed using BLAST and MUSCLE. Sequence types were classified using PubMLST.

**Results:** 2418 swabs from over 1500 patients were screened for *C. difficile*, of which 177 produced *C. difficile* isolates (7%). 179 stool samples were collected during this period, of which over 90% produced isolates. In this dataset, 5 patients transitioned from asymptomatic carriage to CDI, each time without changing sequence type. Additionally, 7 patients transitioned from non-carriage to CDI. While sequencing is not complete (anticipated completion by September), a diverse set of isolates representing over 50 sequence types (ST) was found from 242 sequenced isolates. Of these, only 12 were from ST1 (5%). Strains from ST1 and related STs were more likely to be found in CDI than other strains, and atoxicigenic strains less likely. **Conclusions:** Strains incoming to the hospital are highly diverse and represent much of the genetic diversity within *C. difficile*. ST1 does not represent the predominant strain in our samples, even though it is still more strongly linked to disease than other STs. We find that, in all cases where a patient is asymptomatically colonized before CDI onset, the same strain was isolated before and during CDI. Even with a small sample size, this raises the possibility of being able to identify a subpopulation of patients at greater risk for developing CDI and adjusting medical care appropriately.
Session: **Poster Session B**  
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
Poster Board #: 110  
Title: *Metagenomic Strain Detection with Rainbow Sketching*  
Author: **R. Bovee, C. Smith, N. Greenfield**; One Codex, San Francisco, CA.  

Identifying specific strains and mixtures of strains in complex metagenomic samples is a key challenge in both epidemiology and environmental microbiology. Even sensitive, *k*-mer-based metagenomic classification tools struggle with strain identification due to issues including database quality, contamination, and genetic recombination and other shared homology. In contrast, recent MinHash methods sketch the entire sample (which is inappropriate for complex mixtures) or require a comparison for each available reference (limiting scalability).

We present Rainbow Sketching, an approach that leverages both *k*-mer-based taxonomic classification and MinHash sketching for strain tracking and mixture modeling. Rainbow Sketching first performs taxonomic classification of each individual *k*-mer (“coloring” each *k*-mer) and uses these colors to build a rainbow of discrete sketches. Each taxa-specific sketch may then be compared against a subset of relevant reference genomes - identifying present strains and determining strain-reference novelty. Count data within these sketches also provides a foundation for modeling strain mixtures.

We employed this method in the recent PrecisionFDA CFSAN Pathogen Detection Challenge - achieving the highest overall score detecting *Salmonella* strains against a metagenomic background. We present results from this challenge, several clinical and other real-world datasets, and simulated data to demonstrate the sensitivity and specificity of this approach.
Development of a Serotyping Pipeline Using Whole Genome Sequencing (WGS) for Shigella Identification

Title: Development of a Serotyping Pipeline Using Whole Genome Sequencing (WGS) for Shigella Identification

Author: Y. Wu, H. Lau, T. Lee, D. K. Lau;

Block: FDA, Alameda, CA.

Abstract:
The bacteria Shigella spp. of 4 species and >50 serotypes cause shigellosis, a disease that leads to significant morbidity, mortality, and economic loss worldwide. An estimated 500,000 annual shigellosis cases occur in the US, and the number of cases has been on the rise. Shigellosis is transmitted through the fecal-oral route, and about one-third of these cases are foodborne. Serotyping (speciation) is an important tool for Epidemiological surveillance that informs future policy making for outbreak control and vaccine development.

Classical Shigella serotyping based on serology is tedious, time-consuming, limited by the availability of sensitive and serotype-specific antibodies, and its interpretation often interfered by cross-reactivity. Modern molecular diagnostic assays are fast and sensitive but does not distinguish Shigella at species level or even from the closed related enteroinvasive Escherichia coli (EIEC) strains. Due to its high discriminating power, whole genome sequencing (WGS) holds the promise to replace the conventional Shigella serotyping with a faster and more accurate in silico serotyping. However, analysts trained as Laboratory Microbiologists do not usually possess sophisticated bioinformatics skills. Some serotypes of Shigella are determined by both O-antigen biosynthetic genes and O-antigen modification enzymes, which can be complicated to interpret. We have developed an automated workflow that utilizes limited computational resources to accurately and rapidly determine Shigella serotypes using WGS data from Shigella and EIEC strains available in the laboratory and on NCBI SRA. To conserve time and computational resources, raw WGS reads are subjected to alignment with an in-house curated reference sequence database composed of Shigella serotype determinants and genus- and species-specific sequences as indicators to exclude non-Shigella isolates. Serotype prediction is made based on sequence hits that pass threshold levels of coverage and accuracy. Operators with minimal computer programming skills and knowledge in Shigella genetics can obtain an unambiguous interpretation using this pipeline. For pair-ended fastq reads of < 1.7 GB, the turn-around time is under 5 minutes. This pipeline will be further optimized and streamlined for accuracy, ease of use, and confidence of predictive values before validation. We are also expanding the reference sequence database by constantly updating it with newly available sequences from provisional serotypes. This pipeline is the first step towards building a comprehensive WGS-based analysis pipeline of Shigella spp. for outbreak investigation and control in a field laboratory setting, where speed is essential.
Quinolone Resistance Mechanisms Found in *E. coli* from Four Animal Species in Norway

H. Kaspersen¹, C. Sekse¹, J. S. Slettemeås¹, R. Simm², M. Norström¹, H. Sørum³, A. Urdahl¹, K. Lagesen¹;

¹Norwegian Veterinary Institute, Oslo, NORWAY, ²Department of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, NORWAY, ³Institute of Food Safety and Infection Biology, Norwegian University of Life Sciences, Oslo, NORWAY.

Quinolones and fluoroquinolones are regarded as critically important for human health, but increased use of these compounds have been linked to increased occurrence of resistance. In Norway, fluoroquinolones are used in negligent amounts in livestock, and prophylactic use is prohibited. Nevertheless, low levels of quinolone resistant *E. coli* (QREC) have been observed in a high proportion of the samples analysed in the monitoring programme for antimicrobial resistance in the veterinary and food production sectors (NORM-VET). To better understand the occurrence of QREC, the resistance mechanisms present in selected isolates from the NORM-VET programme are characterized. *E. coli* isolates were defined as QREC when they grew in the presence of ciprofloxacin and/or nalidixic acid at concentrations above the epidemiological cut-off values of 0.06 µg/ml and 16 µg/ml, respectively. QREC isolates were randomly selected and grouped based on animal species of origin, minimum inhibitory concentration (MIC) for ciprofloxacin and nalidixic acid, and the number of additional resistant phenotypes, resulting in 285 isolates. The MIC ranges of the isolates for ciprofloxacin and nalidixic acid were 0.03 - 16 µg/ml and 4 - 256 µg/ml, respectively. Whole genome sequencing on Illumina HiSeq2/3/4000 with Nextera Flex/XT library prep was performed on the isolates. The resulting sequences were run through the Bifrost pipeline (github.com/NorwegianVeterinaryInstitute/Bifrost) for quality control, antimicrobial resistance gene identification, and multilocus sequence typing (MLST). Acquired resistance genes and mutations in intrinsic genes are identified from reads by mapping to a reference database, followed by local assemblies. Preliminary results suggest that over 80 % of the isolates have at least one mutation in the *gyrA* gene, less than 30 % in the *gyrB*, less than 50 % for *parC* and above 60 % for *parE*. Further analysis is being done and results will be presented.
Introduction: Sharing of *Pseudomonas aeruginosa* (Pa) strains between cystic fibrosis (CF) patients with chronic infection is relatively common. It is unclear how frequent Pa strain sharing is in new-onset infections occurring earlier in CF, when infections are treated with antibiotic eradication therapy (AET) and epidemic strains infrequently encountered. We sequenced Pa isolated from sputum of children prior to initiation of inhaled AET, to determine the frequency of mixed strain infection, strain sharing, and their association with AET failure.

Methods: We sequenced 342 Pa isolates using Illumina technology, collected from 65 children with 75 distinct episodes of new-onset infection (episodes at least 1 year apart, AET failure in 27% of episodes) between 2012 and 2016. Up to 10 isolates were sequenced per episode. We performed first-pass analysis of population structure by building phylogenies with 1) Assembly and Alignment Free (AAF), 2) a pairwise distance matrix generated from assemblies using Mash and 3) a conventional mapping step and SNP alignment. We further investigated clusters suggestive of strain sharing by mapping genomes from each cluster to closely related references, using 3 pipelines; 1) Bacteria and Archaea Genome Analyser (BAGA), 2) Snippy and 3) an in-house pipeline. Maximum likelihood phylogenetic trees were generated from Single Nucleotide Polymorphisms (SNP) alignments with IQ-TREE. Strain sharing, which could result from direct/indirect transmission or a common environmental reservoir, was inferred based on detection of appropriate topological signal in these trees (strains from different patients exhibiting close monophyletic, or paraphyletic, relationships). Univariate logistic regressions were used to assess associations between mixed infection, strain sharing and AET failure. All statistical analyses were done using SAS 9.04.01. Results: Pairwise SNP differences between closely related isolates differed depending on the pipeline used and how closely related the reference, but tree topologies were broadly similar. A large number of patients shared Pa strains with other patients (N=25/65, 40%). Mixed infection (two or more strains present in sputum concurrently) occurred in 12/75 episodes (16%). Having a mixed infection was significantly associated with sharing of Pa strains (unadjusted OR 10.7, 95% CI 2.2; 53.7, p <0.01) but was not associated with AET failure. Furthermore, strain sharing was not associated with AET failure. Conclusions: A large proportion of patients were infected with a Pa strain shared with other patients; the reason for this requires for further investigation. Mixed lineage Pa infections were relatively frequently observed in new-onset episodes and were associated with strain sharing between patients. Tree topologies for individual clusters were similar regardless of SNP calling pipeline used despite variation in pairwise SNP difference between isolates.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 114  
**Title:** Antimicrobial Resistance Prediction by Whole Genome Sequencing in MRSA and VRE: A Real-World Application

**Author Block:** A. Babiker, M. M. Mustapha, K. A. Shutt, C. D. Ezeonwuka, S. L. Ohm, M. P. Pacey, J. Marsh, V. S. Cooper, Y. Doi, L. H. Harrison; University of Pittsburgh, Pittsburgh, PA.

**Background:** The antimicrobial resistance (AMR) crisis represents a serious threat to public health and the healthcare economy and has resulted in concentrated efforts to increase development of rapid molecular diagnostics for AMR. In combination with publicly-available web-based AMR databases, whole genome sequencing (WGS) offers the capacity for rapid detection of antibiotic resistance genes. Here we studied the concordance between WGS-based resistance prediction and phenotypic susceptibility testing results for methicillin-resistant Staphylococcus aureus (MRSA) and vancomycin resistant Enterococcus (VRE) clinical isolates using publicly-available tools. **Methods:** Clinical isolates prospectively collected at the University of Pittsburgh Medical Center between December 2016 and December 2017 underwent WGS. Antibiotic-resistant gene content was assessed from assembled genomes by BLASTn search of online databases ResFinder and the Comprehensive Antibiotic Resistance Database (CARD). Concordance between WGS-predicted and phenotypic susceptibility as well as sensitivity, specificity, positive and negative predictive values (NPV, PPV) were calculated for each antibiotic/organism combination, using the phenotypic results as the gold standard. **Results:** Phenotypic susceptibility testing and WGS results were available for 109 and 105 MRSA and VRE isolates respectively. Out of a total of 1,058 isolate/antibiotic combinations overall concordance was 98.8% with a sensitivity, specificity, PPV, NPV of 98.0% (95% CI, 0.97-0.99), 99.1% (95 % CI, 0.98-0.99), 98.5% (95% CI, 0.97-0.99), 99.0% (95% CI, 0.98-0.99), respectively. Identification of point mutations in housekeeping genes increased the concordance to 99.3% and the sensitivity to 99.5% (95% CI, 0.98-0.99) and NPV to 99.8% (95% CI, 0.99-0.99). **Conclusion:** WGS can be used as a reliable predictor of phenotypic resistance for both MRSA and VRE using readily available online tools.
Deep Sequencing of a Measles Vaccine Strain Reveals Complexity of Defective Interfering Genomes

Defective interfering particles (DI) of measles virus (MeV) frequently arise in cell culture, suppressing the replication of standard virus. Paramyxovirus DIs are immunostimulatory, so elucidation of the formation and function of DIs is important to more fully understand MeV replication. Many of the complex truncation jump points of DI RNAs arise from a “copy-back” mechanism involving disassociation of the polymerase complex and reattachment in opposing strand-orientation. DIs are challenging to quantify with traditional molecular techniques. We serially passaged the MeV vaccine strain, Moraten, in Vero-hSLAM and MRC5 cells, using conditions that promote DI formation. A MIQE-standard RT-qPCR assay using SYBR chemistry was developed to measure ratios of MeV full-length genomic and DI RNA species and to minimize experimental noise arising from viral mRNAs. RT-qPCR data were validated by a two-step, end-point PCR procedure that detects a copy-back polarity switch in the DI RNA sequence. RNA extracted from DI-containing cell lysates were sequenced using stranded Illumina chemistry and analyzed for DI jump points by various in silico methods. Gapped read alignments were used in conjunction with R/Bioconductor ranged data processing to detect the true diversity of DI content in the samples. A high diversity of truncation jump points was observed and stranded sequencing data suggested replication of DI genomes. Cyclic suppression of standard virus titers was observed along the passage series, and was inversely correlated with concentrations of DI RNAs determined by RT-qPCR. DIs were detected after serial passage in Vero-hSLAM cells but not MRC5 cells. The findings represent novel evidence of DI complexity in a laboratory passaged MeV vaccine strain; DIs were stably and independently observed in replicate trials over 20 passages. These findings suggest that cell-specific mechanisms affect DI formation, and that NGS methods are of utility for the discovery of novel RNA populations in paramyxovirus-infected cells.
Abstract: The generation of genomic data from microorganisms has revolutionized our abilities to understand their biology, but it is still challenging to quickly and cheaply obtain the complete genome sequence of microbes in an automated, high-throughput manner. While the advent of second-generation sequencing technologies provided significantly higher data throughput, their shorter read lengths and more pronounced sequence-context bias led to a shift towards resequencing applications. Recently, single molecule real-time (SMRT) DNA sequencing has been used to generate sequencing reads that are much longer than second-generation or even Sanger sequencing reads, facilitating de novo genome assembly and genome finishing. Here we tried to develop a novel multiplex strategy to make full use of the capacity and characteristics of SMRT sequencing in microbe genome assembly. We first used error-free simulations to evaluate the practicability of assembling SMRT genomic sequencing data from multiple microbes into finished genomes once at a time. And then we compared the influence of some key factors, including sequencing coverage and read length, on multiplex assembling. Our results showed that long-read genomic sequencing inherently provided the ability to assemble genomic sequencing data from multiple microbes into finished genomes due to its long read length. This approach might be helpful for the various groups of microbial genome projects or metagenomics research.
Antibiotic resistance is a growing health crisis in the US accounting for over 20,000 deaths annually. Environmental bacteria can act as reservoir of opportunistic pathogens despite a lack of exposure. Resistant microbes may have a significant negative impact on the health of Alaskans. Identifying specific antibiotic resistant microbes is essential for quick and appropriate treatment. Here we demonstrate a rapid, automated, and portable sequencing platform. In this proof of concept, each library was constructed from a single cultured microbial isolate using the VOLTRAX, a rapid, automated library preparation. Sequencing was carried out using the portable Nanopore DNA sequencer, MinION. We demonstrate the results of our long read bioinformatic pipeline for assembly, contig polishing, and annotation. All of these steps were carried out by two undergraduates with introductory laboratory skills. Importantly, these methods allow us to better understand the environmental reservoir of antibiotic resistance in Alaska. These results demonstrate the potential application of the portable library preparation and sequencing for a mobile biosurveillance laboratory.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 118
Title: Metagenomics for Diagnosis of Sterile Site Infection: Balancing Automation with Expert Interpretation
Author: C. Anscombe, A. Nguyen, N. Le, H. Nguyen, P. Ashton, T. Le; Oxford University Clinical Research Unit (OUCRU), Ho Chi Minh City, VIET NAM.

Background: A quick literature search will demonstrate the increasing popularity of metagenomic sequencing methods to diagnose patients with suspected infections when other methods have failed. However, there is a need to carry out these analyses on larger, prospective cohorts, to determine sensitivity and specificity. At OUCRU, Vietnam, we are investigating its use in central nervous system (CNS) infections and in patients with fever of unknown aetiology. In order to examine this data effectively we have developed an analysis pipeline which is rapid, requires low RAM, and is designed for clinicians or microbiologists to use. Methods: The pipeline takes raw reads from Illumina sequencing, removes host reads and classifies the remaining samples using CLARK in light mode. After classification, a prediction of genome coverage is made for each organism identified based on number of reads and the genome size of the organism. If a threshold is met, the reference for that taxon ID is downloaded and sample reads mapped. Outputs include mapping statistics such as genome coverage and number of reads mapped. A report on the frequency at which taxon IDs are found across the run is automatically generated, allowing users to consider contamination. Users can customize the classification database to fit with need, define the relevant host genome, input contamination libraries and specify taxa to ignore in analysis based on local knowledge. Results: The pipeline was used to analyze metagenomic sequencing results from 71 CSF samples collected from patients presenting with CNS infection in Vietnam. After pipeline completion, the number of reference mapping analyses was 104. Different sub-types of torque teno virus accounted for 33 of these, and were removed from the analysis. The results were then edited for clinical significance by a microbiologist. Results identified pathogens in 17 samples; 8 Streptococcus suis, 4 enteroviruses, two cases of mumps and one S. pneumoniae, Japanese encephalitis and Varicella-zoster virus (VZV). In addition, Hepatitis B was identified in 5 cases, but was not considered a cause of CNS disease, but merely reflective of the high incidence of Hepatitis B in Vietnam. Genome coverage of these pathogens varied from 0.83% to 81.33%. All findings were confirmed with specific PCR, with Ct values ranging from 27 to 40. Use of an arbitrary cut off in reference genome coverage led to missing VZV and produced 2 false positives (a polytropic provirus and Streptococcus agalactiae ), which were all negative by PCR. Showing that just as in a culture based diagnostic approach there is no replacement for expert interpretation of results. Conclusions: Bioinformatics can help to automate processing of metagenomic sequencing but, interpretation remains the domain of human intelligence. Building bioinformatics tools with this in mind will enable more rapid uptake of metagenomics for diagnosis from sterile sites.
Microorganisms are the most diverse and abundant life forms on Earth and account for a large portion of the Earth’s biomass and biodiversity. To date though, our knowledge regarding microbial life is lacking, as it is based mainly on information from cultivated organisms. Indeed, microbiologists have borrowed from astrophysics and termed the ‘uncultured microbial majority’ as ‘microbial dark matter’. The realization of how diverse and unexplored microorganisms are, actually stems from recent advances in molecular biology, and in particular from novel methods for sequencing microbial small subunit ribosomal RNA genes directly from environmental samples termed next generation sequencing (NGS). This has led us to use NGS that generates several gigabases of sequencing data in a single experimental run, to identify and classify environmental samples of microorganisms. In metagenomics sequencing analysis (both 16S and shotgun), sequences are compared to reference databases that contain only small part of the existing microorganisms and therefore their taxonomy assignment may reveal groups of unknown microorganisms or origins. These unknowns, or the ‘microbial sequences dark matter’, are usually ignored in spite of their great importance. The goal of this work was to develop an improved bioinformatics method that enables more complete analyses of the microbial communities in numerous environments. Therefore, NGS was used to identify previously unknown microorganisms from three different environments (industrials wastewater, Negev Desert’s rocks and water wells at the Arava valley). 16S rRNA gene metagenome analysis of the microorganisms from those three environments produce about ~4 million reads for 75 samples. Between 0.1-12% of the sequences in each sample were tagged as ‘Unassigned’. Employing relatively simple methodology for resequencing of original gDNA samples through Sanger or MiSeq Illumina with specific primers, this study demonstrates that the mysterious ‘Unassigned’ group apparently contains sequences of candidate phyla. Those unknown sequences can be located on a phylogenetic tree and thus provide a better understanding of the ‘sequences dark matter’ and its role in the research of microbial communities and diversity. Studying this ‘dark matter’ will extend the existing databases and could reveal the hidden potential of the ‘microbial dark matter’.
Whole-Genome Sequencing of Zika Virus Directly from Clinical Samples

K. Kamelian¹, A. Olmstead², V. Montoya², W. Dong², M. Morshed³, P. R. Harrigan¹, J. Joy²; ¹University of British Columbia, Vancouver, BC, CANADA, ²BC Centre for Excellence in HIV/AIDS, Vancouver, BC, CANADA, ³BC Centre for Disease Control, Vancouver, BC, CANADA.

Background: In 2016, The World Health Organization declared the Zika virus a Public Health Emergency of International Concern due to the increasing prevalence of Zika virus infections in the Americas. The Zika virus has been associated with increased incidence of the neurological condition Guillian-Barré syndrome and birth defect microcephaly. Routine surveillance tools currently rely on PCR amplification, Sanger sequencing, and antibody-based tests to identify new cases of Zika infections. However, whole-genome sequencing (WGS) of the Zika virus may present certain advantages over other surveillance tools by providing more detailed information on viral phylogenetic clustering, transmission, and geography. Methods: Specimens from five subjects with travel-acquired Zika virus infection (putatively from Belize, Mexico, an undisclosed Caribbean region, Barbados, and Panama) were obtained from the British Columbia Centre for Disease Control Public Health Laboratory (BCCDC) and had a range of cyclic threshold (Ct) values (21 - 33). WGS of Zika virus was performed on an Illumina MiSeq using a previously published procedure designed to overcome some of the limitations of low viral load and partially degraded samples by amplifying several short amplicons to create a tiling path across the Zika virus genome (Quick et al., 2017). Sequences were analyzed for depth of coverage and total number of reads including total quality trimmed reads, viral reads, and human reads. Phylogenetic analysis was performed to investigate geographic clustering of travel-related cases. Results: Consensus sequences ranging from 8 - 10.5 kb were obtained for the five samples. Higher Ct values were correlated with lower coverage, lower number of viral reads, higher number of human reads, and overall lower depth. Median depth of coverage of the samples was 24,000 (IQR: 17,000-25,000). Although some contigs had low depth of coverage (less than 10 reads), they still provided adequate genome coverage for the regions sequenced. Phylogenetic analysis of sequencing data confirmed the suspected regions of Zika infection for two of five samples. Three samples were missing reference genomes of suspected areas of infection. However, they clustered within close geographical proximity to neighboring regions. Conclusions: Our results highlight the usefulness of WGS using the tiling amplicon method in a clinical setting. WGS of the Zika virus allows insight into the origins of infection, transmission patterns, and the genetic diversity of travel related cases. However, our samples gave rise to five sources of the Zika virus infection suggesting that the complexity and global movement of the Zika virus epidemic is likely to limit precise interpretations of the origin of travel related cases and is dependent on availability of reference sequences from regions of interest.
Session: Poster Session B  
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
Poster Board #: 121  
Title: Identifying Putative Transmission Clusters of The Multidrug-Resistant *E. coli* ST131-H30 Lineage Among U.S. Children Using Whole Genome Sequencing  
Author: A. Miles-Jay¹, S. J. Weissman², A. L. Adler², J. G. Baseman¹, D. M. Zerr²;  
Block: ¹University of Washington, Seattle, WA, ²Seattle Children’s Hospital, Seattle, WA.

**Introduction:** *E. coli* ST131-H30 is a globally disseminated lineage that is implicated in rising rates of multidrug resistance among extraintestinal *E. coli* infections. Despite the public health significance of this pathogen, its transmission dynamics are poorly understood. This is in part due to ST131-H30’s capacity for prolonged subclinical intestinal colonization, likely resulting in a plethora of “silent” transmission events that are difficult to capture directly. We assessed the ability to detect putative transmission clusters among *E. coli* ST131-H30 isolates collected from U.S. children during routine clinical care. **Methods:** We applied whole genome sequencing and a novel framework for transmission cluster detection to clinical *E. coli* ST131-H30 isolates collected in a multicenter surveillance study that took place from 2009-2013 at 4 geographically diverse U.S. children’s hospitals. Isolates were sequenced on an Illumina NextSeq platform. Quality filtered and trimmed sequencing reads were mapped to a high-quality ST131-H30 reference genome and core genome single nucleotide variants were identified. The R package `transcluster`—which probabilistically infers the number of transmission events separating cases using pairwise genomic distance and sampling dates—was used to identify and characterize putative transmission clusters where the implied number of transmissions was less than 25 with a probability of 80% (the default settings). **Results:** A total of 126 *E. coli* ST131-H30 isolates were included. Twelve isolates (9.5%) were placed into 6 putative transmission clusters; each cluster contained 2 isolates and no clusters spanned multiple study sites. The time between sampling in a cluster ranged from 1 to 199 days. Five of the 6 clusters were composed of the CTX-M-15-type extended-spectrum beta-lactamase-producing subclone of ST131-H30; 1 cluster was composed of non-ESBL producing H30 isolates. The implied number of transmission events separating isolates in a single cluster ranged from 1-18 events. The clusters contained a mix of hospital associated (n = 5), healthcare-associated community onset (n = 3), and community-associated (n = 4) infections. Two instances of plausible nosocomial transmission were identified. **Conclusions:** The integration of whole genome sequencing data and a novel framework for transmission cluster detection revealed putative transmission clusters among *E. coli* ST131-H30 isolates collected during routine clinical care. Although geographic location was not explicitly incorporated into the analysis, all clusters sorted by geographic site, strengthening their epidemiologic plausibility. Whole genome sequencing of clinical isolates could guide more detailed and resource-intensive sampling efforts designed to elucidate transmission pathways of difficult-to-track and worrisome lineages like *E. coli* ST131-H30.
Virulence Characteristics and an Action Mode of Antibiotic Resistance in Multidrug-Resistant *Pseudomonas aeruginosa*

W. Hwang;
Department of Microbiology and Immunology, Yonsei University, Seoul, KOREA, REPUBLIC OF.

*Pseudomonas aeruginosa* displays intrinsic resistance to many antibiotics and known to acquire actively genetic mutations for further resistance. In this study, we attempted to understand genomic and transcriptomic landscapes of *P. aeruginosa* clinical isolates that are highly resistant to multiple antibiotics. We also aimed to reveal a mode of antibiotic resistance by elucidating transcriptional response of genes conferring antibiotic resistance. To this end, we sequenced the whole genomes and profiled genome-wide RNA transcripts of three different multi-drug resistant (MDR) clinical isolates that are phylogenetically distant from one another. Multi-layered genome comparisons with genomes of antibiotic-susceptible *P. aeruginosa* strains and 70 other antibiotic-resistance strains revealed both well-characterized conserved gene mutations and distinct distribution of antibiotic-resistant genes (ARGs) among strains. Transcriptions of genes involved in quorum sensing and type VI secretion systems were invariably downregulated in the MDR strains. Virulence-associated phenotypes were further examined and results indicate that our MDR strains are clearly avirulent. Transcriptions of 64 genes, logically selected to be related with antibiotic resistance in MDR strains, were active under normal growth conditions and remained unchanged during antibiotic treatment. These results propose that antibiotic resistance is achieved by a “proactive” response scheme, where ARGs are constitutively expressed even in the absence of antibiotic stress, rather than a “reactive” response. Bacterial responses explored at the transcriptomic level in conjunction with their genome repertoires provided novel insights into (i) the virulence-associated phenotypes and (ii) a mode of antibiotic resistance in MDR *P. aeruginosa* strains.
**Abstract:**

Microorganisms are ubiquitous and important to the proper functioning of ecosystems; including aquatic milieu, where they play vital roles in the water cycles and removal of nutrients and toxins. Hence, studying these microbes are very essential. Prior to now, non-culturable microbes are difficult to study, however, the advent of metagenomics analysis has helped to solve this problem. In this present study, the gene profile of microbial compositions of used water from animal research farm was analysed, in order to obtain a scanned profile of all resident microbiome in the water. To analyse the microbial communities as depicted in the sequenced data, MetaPhlAn2 was used. The reads were analysed and combined to form a merged abundance table. This table was edited and viewed using LibreOffice Calc. A heatmap showing the abundance profiles of the microbes was generated using Hclust2. A cladogram showing taxonomic relatedness was captured using GraphlAn. This was done by rendering trees and annotating them with microbial names and relative abundances. Pie-charts, showing specific comparisons based on various clade, were generated using Krona. The analysis of the microbial samples showed that the environment was dominated by Bacteria (99.88%). The sample also showed that Archaea (0.07%) and Viruses (0.05%) were present, in very small populations. Upon further analysis, it was shown that about 5 phyla were present within the bacterial population namely: Actinobacteria (0.5%), Bacteroidetes (9%), Firmicutes (3.3%), Proteobacteria (86.6%) and Spirochaetes (0.6%). A total of 40 different genera were identified with the genera Thauera making up 74% of the entire population. This comes as no surprise as thauera is a denitrifying bacteria playing a crucial role in the waste water ecosystem. Thauera plays an important role in the removal of nitrogen nitrate and other aromatic compounds. For this reason, Thauera is usually detected in most wastewater treatment samples. Another notable genera worth noting among the bacterial population is the genera Thiomonas making about 5% of the bacterial population. The viruses present in the effluent are composed majorly of the genus Siphoviridae and Gammaretrovirus. The Kingdom Archaea was found to consist of only the genera Methanobrevibacter. Metagenomic analysis provide insight into the microbial community present in the waste water effluent. We were able to analyse the various microorganisms present as well as their relative abundances. We were also able to use various tools to provide graphical illustrations that aided our analysis. The results revealed that a large percentage of the microorganisms present were bacteria and we were able to view their diversity. A huge part of this bacterial presence was directly involved in the wastewater ecology and had major roles in the breakdown of chemical compounds present in the water.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 124  
**Title:** Nanopore Sequencing for AMR Detection and Characterization  
**Author Block:** Y. Fan, W. Timp, T. Simner, P. Tamma, Y. Bergman;  
1Johns Hopkins University, Baltimore, MD, 2Johns Hopkins Hospital, Baltimore, MD.

**Abstract:**

**Background:** The continuing threat of antimicrobial resistance poses an urgent global public health concern. Early and accurate detection of resistance mechanisms can both prevent the dissemination of resistant organisms within the healthcare environment and ensure patients are placed on early and effective antibiotic therapy. To this end, we leveraged the long reads, low overhead, and real time analysis capabilities of nanopore sequencing in order to detect both acquired resistance genes and chromosomal mutations that potentially confer antimicrobial resistance. **Methods:** Forty clinical *Klebsiella pneumoniae* isolates with a variety of resistance mechanisms from patients hospitalized at The Johns Hopkins Hospital were sequenced on the Oxford MinION platform. Genomes were assembled using canu, and corrected with signal level algorithms implemented in the nanopolish software package. These isolates were also sequenced on the Illumina platform, and the more accurate short read data were used to further correct the assemblies. Abricate was used to screen the contigs for resistance genes, using several databases, including CARD, Resfinder, and PlasmidFinder. Chromosomal mutations and their consequences in the amino acid domain were identified using custom C++ code. **Results:** We found that disagreements between the nanopolished and Illumina polished assemblies clustered near methylation motifs. By examining these errors, and by using improved signal models for these motifs in amethylationaware version of nanopolish, we can build an assembly using only nanopore data that achieves ~99.8% identity with illumina polished assemblies. We are continuing to examine the locations and motifs of the remaining ~0.2% of errors to identify new and better approaches to polish nanopore-only assemblies. This remaining ~0.2% is important because it makes accurate prediction of protein translation, and hence truncating or missense mutations difficult to detect. Corrected assemblies allowed us to identify a variety of small mutations noted in the literature to be responsible for resistance phenotypes. By limiting our analysis pipeline to 52,000 reads with an average length of 10 kb, we are able to sequence and build a high quality genome in under 8 hours using a machine with 36 cores and 72 GB RAM. **Conclusions:** We are developing tools that apply nanopore sequencing for rapid and accurate identification of antibiotic resistance mutations, which will clinicians in placing critically-ill patients on early and effective antibiotic therapy. As we collect and sequence more isolates, and accrue information about genetic features that give rise to antimicrobial resistance, we will increase the utility of real time sequencing assays for diagnostic purposes.
ASA³P: An Automatic and Highly Scalable Pipeline for Bacterial Genome Assembly, Annotation and Higher-level Analyses

O. Schwengers¹, A. Hoek¹, M. Schneider¹, M. Fritzenwanker², L. Falgenhauer², J. Falgenhauer², T. Hain², T. Chakraborty², A. Goesmann¹;
¹Bioinformatics and Systems Biology, Justus-Liebig University Giessen, Giessen, GERMANY, ²Institute of Medical Microbiology, Justus-Liebig University Giessen, Giessen, GERMANY.

Background: Major technological advances and the dramatic decrease in costs of bacterial whole genome sequencing is having an unprecedented effect in genome epidemiology and metagenomics. These exciting developments require the establishment of effective, efficient and scalable bioinformatics software tools for data processing and analysis of the high throughput data obtained before scientific interpretation can take place. Methods: In order to solve core bioinformatics tasks such as quality trimming, assembly and annotation, ASA³P takes advantage of published and well performing third party tools and combines them with comprehensive databases. It is a modular software pipeline comprising a core application implemented in Java and Groovy together with cluster distributable scripts implemented in Groovy. HTML reports take advantage of modern and interactive JavaScript libraries. For massive scalability our pipeline integrates well with Sun Grid Engine compatible compute clusters. Within cloud computing environments the software is able to setup complex hardware and software infrastructures and thus is able to automatically create its own compute clusters. Results: Here, we introduce ASA³P, a fully automatic and scalable assembly, annotation and higher-level analysis pipeline for bacterial genomes. The pipeline conducts all of the necessary data processing steps, i.e. quality clipping and assembly of sequencing reads, scaffolding subsequent contigs and annotation of genome sequences. Furthermore, ASA³P performs comprehensive genome characterizations and analyses, e.g. for taxonomic classification, and detection of both AMR genes and virulence factors. Results are presented via an HTML based user interface providing aggregated information, interactive visualizations and access to intermediate results in standard bioinformatic file formats. ASA³P is available in two versions: a local Docker container for small-scale projects and an OpenStack cloud version able to automatically create and manage its own self-scaling compute cluster. Discussion: ASA³P is a software tool enabling the automatic processing, assembly, annotation and higher level analysis of bacterial NGS whole genome data in a comfortable but high-throughput manner. The burden of technical complexity is overcome by simple setup routines and the use of Docker and OpenStack images. Thus, automatic and standardized analysis of hundreds of bacterial genomes is now feasible on a daily basis.
Abstract:

Soil microbiome responses to changing environmental conditions are manifested as shifts in community structure and/or modification of activity. However, molecular-level details underlying functional responses of soil microorganisms to perturbation are largely unknown. Here, we demonstrate a multi-omics approach to determine the impact of environmental perturbations on the soil microbiome across taxonomic and functional levels. Kansas native prairie soil samples from three field locations were either treated with glycine as a model root exudate, or perturbed by changing moisture conditions. The microbiome response was assessed using a suite of omics measurements: 16S rRNA amplicon sequencing, metagenomics, metatranscriptomics, and metabolomics. The soil microbiome responded to glycine at the functional level, but not at the community structure level. In contrast, soil drying shifted both the microbiome composition and function. A major challenge in soil microbial ecology is the extraordinary phylogenetic and functional diversity of the soil microbiome in association with the physico-chemical complexity of the soil habitat. Here by using a multi-omics approach, we elucidated the phenotypic response of the soil microbiome across different levels of expression; thus providing a proof-of-concept for use of this approach to assess key physiological traits expressed by the soil microbiome.
Microbial Diversity of New Orleans Groundwater Using Illumina Miseq

Groundwater contamination will result in poor drinking water quality, loss of water supply, and pose human health in great risk. Increased attention has been devoted to the direct detection of pathogenic organisms in groundwater by using next-generation sequencing (NGS). We investigated microbial biodiversity of groundwater using Illumina Miseq. Water samples were collected from 55 private wells in New Orleans Louisiana. Our results indicate twenty bacterial phyla. Proteobacteria was the most dominant phylum in most of samples (relative abundance: 71.1%), followed by Chlorobi (5.1%), Actinobacteria (4.2%), Chloroflexi (3.3%), Cyanobacteria (2.2%), and Bacteroidetes (2.0%) (Fig. S5). At the genus level, five genera were abundant (> 3%) in well water samples with Methylomonas (5.3%), Methylosinus (3.7%), Mycobacterium (3.4%), Dechloromonas (3.3%), and Thiobacillus (3.1%). The relative abundances of the class of Gammaproteobacteria and Actinobacteria were positively associated with qPCR results of Legionella spp. and mycobacteria respectively. However, the regression analysis showed no significance (p > 0.05). Principal coordinates analysis (PCoA) of unweighted UniFrac indicates patterns of bacterial community composition in groundwater reflect sampling locations.
Abstract: Fermented beverages represent a growing multibillion dollar industry. This is not limited to beer and wine – other products like hard cider and kombucha are among the fastest growing sectors in the industry. Understanding the microbial communities that generate – or spoil – these products are essential to controlling flavor and consistency. Plummeting sequencing costs and new long read sequencing technologies can enable genomics of individual isolates as well as microbial consortia. To this end, we demonstrate the utility of nanopore sequencing for genome and metagenome sequencing of beer, hard cider, and kombucha fermentations. We present de novo genome assemblies of four isolates: a *Saccharomyces cerevisiae* strain from homebrew, two novel *Saccharomyces* and *Pichia* yeast isolates from hard cider, and a *Gluconacetobacter xylinus* strain from a home kombucha culture. We further attempt to sequence the metagenome of the kombucha culture, and show genome assembly of the most common *Acetobacter* strain is possible, although targeted metagenomes such as those based on 16S rRNA are likely better suited to capturing all members of the community. To assemble the isolate genomes, four de novo genome assemblers, MiniASM, Canu, Flye, and SMARTdenovo, were evaluated at varying genome coverages, with SMARTdenovo performing the best based on number of contigs, number of mismatches, and average coding sequence (CDS) length. Quality of the assemblies can be greatly improved by scaffolding to a reference genome with pyScaf and polishing using Nanopolish, increasing the average CDS length by approximately 33% across all assemblies. Yet, the average CDS length of the polished nanopore assemblies were approximately 75% of their reference strain, which could be improved by complementary Illumina sequencing. Yet, without Illumina data, we were able to place the homebrew *S. cerevisiae* strain in the “Beer 2” clade based on multiple mutations to the *MAL11* gene and a specific W497 mutation in the *FDC1* gene, identify new yeast species from hard cider, assemble a genome from a metagenome experiment, and assemble a complete *G. xylinus* genome from an isolate with a contig longer than the existing reference strain. This demonstrates that quality genome assemblies from fermented beverages using nanopore sequencing are possible. The low cost and ease of use of nanopore technology promises high quality genomic information for future strain breeding or engineering, as well as assessing spoilage for process control. Thus, a workflow of nanopore sequencing coupled with de novo assembly using SMARTdenovo, and optionally pyScaf and Nanopolish, can facilitate quality genomics of microbes from fermented beverages and therefore has great potential utility for producers of fermented beverages for the control of flavor and consistency.
Title: Whole Genome Analysis of Clinical Multidrug-resistant *Acinetobacter baumannii* Strains in Vietnam Hospital

**Author Block:**

N. Si-Tuan¹, C. Nguyen², H. Nguyen Thuy³;

¹Thongnhat General Hospital of Dongnai Province, 234 highway 1A, Tan Bien ward, Bien Hoa City, VIET NAM, ²Department of Bioinformatics and Medical Statistics, Vinmec Research Institute, Hanoi, Vietnam, Hanoi, Vietnam, VIET NAM, ³Department of Biotechnology, Faculty of Chemical Engineering, Ho Chi Minh City University of Technology, HCM National University, Ho Chi Minh City, VIET NAM.

**Background:** *Acinetobacter baumannii* is an important nosocomial pathogen that can develop multidrug resistance. In this study, we sought to explore the genomic properties, phylogenetic relationships, and comparative genomics of this pathogen through strain DMS06669 and DMS06670 (isolated from the sputum of two male patients with hospital-acquired pneumonia).

**Methods:** Whole genome analysis of *A. baumannii* DMS0669 and DMS06670 included de novo assembly; gene prediction; functional annotation to public databases; phylogenetic tree construction by average nucleotide identity; pan-genome analysis and antibiotics resistance genes identification. Antibiotics resistance genes *in-vitro* were isolated by PCR and re-confirmed by improved Sanger method. **Results:** The data showed that a total of 19 possible antibiotic resistance genes, conferring resistance to eight distinct classes of antibiotics, were identified in two strains. Nine of these genes have not previously been reported to occur in *A. baumannii*. Comparative analysis of 23 available genomes of *A. baumannii* revealed an open pan-genome consisting of 15,883 genes. All antibiotics resistance genes were isolated. **Conclusions:** Our results provide important information regarding mechanisms that may contribute to antibiotic resistance in the DMS06669 and DMS06670 strain and have implications for treatment of patients infected with *A. baumannii*.

**Keywords:** *Acinetobacter baumannii*, multidrug resistance, pan-genome analysis, comparative whole-genome analysis, next generation sequencing.
Session: **Poster Session B**
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 130
Title: Genomic Characterization of Carbapenem-resistant *Escherichia coli* Isolated from Clinical Samples in Thailand

**Authors:**

K. R. Margulieux¹, A. Srijan¹, S. Ruekit¹, E. Snøsrud², A. Ong², O. Serichanthalergs¹, R. Kormanee³, P. Sukhchat³, A. Jones², P. McGann², J. Crawford¹, B. Swierczewski²;

¹Armed Forces Research Institute of Medical Sciences, Bangkok, THAILAND, ²Walter Reed Army Institute of Research, Silver Spring, MD, ³Queen Sirikit Naval Hospital, Chonburi, THAILAND.

**Abstract:**

**Background:** The spread of multidrug-resistant bacterial pathogens is one of the most dangerous current public health threats, especially in regions such as Southeast Asia with widely unregulated antibiotic usage. Whole genome sequencing (WGS) of clinical isolates can provide additional insights about multidrug-resistant isolates and complement efforts of local clinical laboratories. **Methods:** A total of 183 clinical *Escherichia coli* isolates were collected at Queen Sirikit Naval Hospital in Chonburi, Thailand from October 2016 - January 2018 as part of routine surveillance for multidrug-resistant pathogens. The isolates were verified and underwent antimicrobial susceptibility testing using the BD Phoenix™ 50 and the NMIC/ID 95 panel to screen for carbapenem-resistant isolates. WGS of identified carbapenem-resistant isolates was performed on an Illumina MiSeq Benchtop sequencer and subsequently analyzed. **Results:** Of the 183 multidrug-resistant *E. coli* isolates collected, 168 (91.8%) demonstrated resistance to 3rd generation cephalosporins and 167 (91.3%) to azetronam. Phenotypic resistance to at least one carbapenem tested was observed in 11/183 (6.0%) isolates. 8/11 isolates were positive for carbapenemase production using a CarbaNP test. Genomic characterization of *E. coli* isolates showed that the 11 isolates carried 5 carbapenem-resistance genes between them. Single isolates carried *bla*<sub>NDM-4</sub>, *bla*<sub>KPC-2</sub>, and *bla*<sub>OXA-181</sub>, two isolates carried *bla*<sub>OXA-48</sub>, and six isolates carried *bla*<sub>NDM-5</sub>. The three isolates that tested negative for carbapenemase production with the CarbaNP test carried *bla*<sub>OXA</sub> genes. Isolate relatedness was shown through whole genome comparison. **Conclusions:** A total of 11/183 (6%) multidrug-resistant *E. coli* isolates identified over 16 months at Queen Sirikit Naval Hospital were shown to be carbapenem resistant. Five carbapenem-resistance genes were identified in these 11 isolates. The most prevalent carbapenem-resistance gene was *bla*<sub>NDM-5</sub>, a gene increasingly reported in Southeast Asia. Notably, the isolates carrying *bla*<sub>OXA</sub> genes appeared to have lower phenotypic levels of carbapenemase production compared to other gene types. This may lead to an under-reporting of carbapenem-resistance in the region and treatment complications if not detected during routine clinical screening. It is important to continue long-term surveillance of hospitals in Thailand, and utilizing WGS for in-depth isolate analysis is important to fully understand the current circulation of resistant pathogens and their evolution over time.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 131  
**Title:** All 16S rRNA Gene Variable Regions Essential for Microbiome Survey  
**Author Block:** Q. Yang, C. Franco, W. Zhang; Flinders University, Adelaide, AUSTRALIA.

**Background:** Marine sponges (phylum Porifera) are enriched by host-specific and opportunistic microorganisms that make up to 60% of the mesohyl volume. The majority of these microbes have not been identified. 16S rRNA gene based metagenomics sequencing has become the method of choice to study sponge microbiomes, however, results from amplicon-based analyses that employ only one pair of primers targeting specific variable regions have been found to be grossly under-representative. Therefore, the aims of this study were to test the hypothesis that primers targeting different variable regions of the 16S rRNA gene reveal vastly different parts of the microbiome and to develop an improved approach to reveal a more complete microbial profile. **Methods:** To test the hypothesis, five primers sets covering all the variable 16S rRNA gene regions were validated to reveal the microbiomes of four representative sponge species in different orders on Illumina MiSeq platform. **Results:** A significant increase in microbiome coverage was achieved. 29.5% of phylum-level OTUs and 35.5% class-level OTUs generated from this developed approach could not be recovered by the most commonly used single primer set targeting the V4 region only. In relation to the microbial sequence recovery, this approach could increase the sequence coverage by 93.9 to 549.9% for each of four sponge species when compared to that using the V4 primer set. A further indirect comparison with metagenomics-based microbiome survey demonstrated that the multi-primer approach performed substantially better, especially in revealing unaffiliated taxa, that are either candidate or unassigned. **Conclusions:** Our study indicated that a validated combination of variable region-specific primer sets covering the full length of 16S rRNA gene is essential to analyze sponge microbial communities when using amplicon-based analysis, so as to avoid an incomplete and misleading microbiome profiling. This multi-primer approach can be conveniently applied and represents a fundamental change from conventional single primer set amplicon-based microbiome studies. It could contribute significantly to any microbiome survey projects, to achieve a more comprehensive understanding of the microbial profile. The superior capacity on uncovering the unaffiliated microbial OTUs allows for a greater potential to discover the taxonomic ‘blind spots’ within the largely unknown microbial world.
Factors Associated with Surface Water Microbial Community Structure

T. Chung\textsuperscript{1}, D. L. Weller\textsuperscript{2}, J. Kovac\textsuperscript{1};
\textsuperscript{1}The Pennsylvania State University, State College, PA, \textsuperscript{2}Cornell University, Ithaca, NY.

**Background:** According to the U.S. Geological Survey, surface-water sources accounted for 52\% of all irrigation withdrawals in 2015. However, temporal variation in physical (e.g., turbidity) and chemical (e.g., pH) water quality, and spatiotemporal variation in environmental factors (e.g., weather, proximity to upstream livestock operations) may affect microbial community composition and the microbiological quality of surface water. While a number of studies have investigated drinking water microbiomes, little is known about the microbial communities in surface waters that are used for irrigation. Here we investigated geospatial, weather and landscape factors associated with surface water micro- and mycobiomes.

**Methods:** We characterized the bacterial and fungal community composition of 68 samples collected using Moore swabs from six streams in upstate New York between May and August 2017. Samples were separated into particulate matter (i.e., soil) and water fractions. Total DNA was extracted from fractions using PowerSoil (n=68) or PowerWater (n=46) kits, respectively. Data on physical and chemical water quality, upstream landscape characteristics and weather data were also collected at sampling. Microbial community composition of each sample was determined by Illumina sequencing of PCR-amplified 16S rRNA gene V4 region and ITS sequences. Sequences were processed using Mothur. The resulting OTUs were used to investigate sample biodiversity within and between streams using permutational multivariate analysis of variances (PERMANOVA), clustering, ordination and network analysis.

**Results:** Significant differences in microbial community structure were observed among samples collected from different streams. According to PERMANOVA, these differences may be associated with differences in upstream activity (p<0.05). Three out of 18 samples collected immediately downstream of dairies had a relatively higher abundance of Moraxellaceae or Enterobacteriaceae. Microbial communities also differed between water and soil fractions of individual samples according to the UniFrac-based PCoA clustering, and PERMANOVA (p<0.01). While the dominant families in soil fractions were Rhodocyclaceae, Rubritaleceae, and Sphingomonadaceae; Chitinophagaceae, Enterobacteriaceae, and Moraxellaceae were more abundant in water fractions.

**Conclusions:** Taxonomic composition of soil and water fraction of collected surface water samples differ. Further, the microbial and especially Enterobacteriaceae content in water may be affected by the adjacent land use. Thus, this study provides a baseline data describing surface water microbiome structure that can guide further studies focused on detection of microbiological safety hazards in water.
Whole Genome Sequencing Analysis Reveals That Air-conditioners Cooling Towers are Reservoirs for *Legionella pneumophila* and Lead to Infections with the Same Strains Over Years

**D. Wüthrich**¹, S. Gautsch², P. Brodmann², O. Dubuis³, R. Spieler-Denz⁴, S. Tschudin-Sutter¹, V. Gaia⁵, S. Fuchs⁴, A. Egli¹;

¹University Hospital of Basel, Basel, SWITZERLAND, ²Cantonal Laboratory City of Basel, Basel, SWITZERLAND, ³Viollier, Allschwil, SWITZERLAND, ⁴Medical Services City of Basel, Basel, SWITZERLAND, ⁵National Reference Center for Legionella, Bellinzona, SWITZERLAND.

**Background:** Water supply and air-conditioner cooling towers (ACCT) are known potential sources of *Legionella pneumophila*. Traditional typing methods have low resolution and may not allow reliable identification of transmissions. The advent of whole genome sequencing (WGS) allows high-resolution analysis, and the study of complexity within environmental compartments.

**Materials/methods:** In summer 2017, the health administration of the City of Basel detected an increase of *Legionella pneumophila* infections compared to previous months. An epidemiological and WGS-based microbiological investigation was performed, involving isolates from the local water supply and two ACCTs (n=60), clinical outbreak and non-outbreak related isolates from 2017 (n=8) and those collected between 2003-2016 (n=26). Finally, we also compared the sequenced strains to already published bacterial genomes from 17 countries (n=539).

**Results:** Phylogenetic analysis of the ACCT isolates showed clustering into two groups separated by a few hundred allelic differences. Several strains were found in both ACCTs. Furthermore, we found that isolates from the two ACCTs were highly related to three clinical isolates from 2017. Five clinical isolates from the last decade also found to be closely related to the recent isolates from ACCTs. Finally, we found several clinical isolates to be related to published genomes.

**Conclusions:** Current outbreak-related and historic isolates were linked to ACCTs. ACCTs form a complex environmental habitat in which strains are conserved over years and are exchanged between locations. WGS-based typing allows to explore this complex network, which might have public health implications on the tracing of potential sources and the interpretation of environmental findings.
A recent study using whole genome sequencing (WGS) demonstrated the presence of a clonal outbreak of multidrug resistant (MDR) TB in Mumbai. Thus transmission of Mtb in the community appears to be one of the most significant contributors to the current epidemic like situation in Mumbai and elsewhere, and emerges as a key intervention point for the public health system. While nosocomial transmission and transmission in public spaces have been identified for intervention, household transmission is overlooked. One of the reasons for the neglect is due to limited data. Previous inability to ascertain the path of transmission has been recently overcome by the use of whole genome sequencing (WGS) which has successfully traced several outbreaks of TB. Here we determine the transmission of Mtb in low socio-economic households, in a defined slum cluster and an adjacent slum rehabilitation (SRA) cluster in Mumbai. In these low income settlements the air quality has been found to be poor which can significantly increases the risk of airborne infection. Additionally the spatial adjacency of these slum settlements pose a potential risk to increasing vulnerability. Using WGS of Mtb isolated for TB patients in the two locations, the study demonstrates the proportion of TB caused by household transmission. Using phylogenomics, bayesian estimation of risk of infection and GIS mapping, the study will conclusively trace the transmission chains in the locale. By overlaying this information with modelling of the household built environment, the study proposes to understand the potential contribution of such layouts and the effect of spatial autocorrelation to the increased transmission, the study will device a novel public health tool for Real time monitoring and mapping of TB transmission. Additionally, this study will contribute towards understanding the relationship between TB transmission and slum-household clustering through a spatio-temporal analysis route. This spatial analysis route adds to the novelty of this study.
Abstract: A total of 40,600 reads were obtained from the two 16S rRNA gene samples sent for Illumina sequencing after removal of reads corresponding to Cyanobacteria and other chimeras. In our study, regardless of the plant tissue, members of phylum Firmicutes were the most abundant, followed by those of Proteobacteria, Bacteroidetes, Actinobacteria, Spirochaetes, and Tenericutes. Analysis of the sequences revealed the presence of a core bacterial endomicrobiome comprising mainly of Acinetobacter sp., Heliorestis sp. and Thiomonas sp. in the root and shoot metagenomic samples. The root metagenome of rice, however, showed more abundance and diversity of bacterial endophytes than the shoot sample which include species of Pseudomonas and Stenotrophomonas. All the bacterial endophytes identified are previously reported to exhibit PGP in host plants through several mechanisms, such as phytohormone production, nitrogen fixation, phosphate solubilization, siderophore production, etc. Through the culturable approach, different species of Pseudomonas and Methylobacterium were isolated from the callus cultures of Kalonunia rice cultivar which are reported to have good PGP abilities. Conclusion: Our findings indicate that a typical metagenomic diversity of endophytic bacteria could be predicted through NGS for the development of suitable bacterial consortia using selected endophytic isolates as bioinoculants to improve rice crop productivity.
Abstract: The herpes simplex viruses HSV-1 and -2 are ubiquitous human pathogens responsible for a large burden of disease worldwide, manifesting as oral and genital ulcers, neonatal disease, encephalitis and keratitis. To date, most HSV genomics has been performed on culture isolates, raising concerns that these genomes may not accurately represent the clinical specimens from which they were derived. We have developed and validated an approach that combines a DNA oligonucleotide hybridization panel with a bioinformatic pipeline that allows the recovery of near-complete HSV genomes directly from clinical specimens with Illumina sequencing. Our computational pipeline performs rapid assembly and annotation of whole viral genomes starting from raw reads, allowing the recovery of near-full-length genomes from specimens with as low as $10^2$ HSV copies/ml and 100,000 reads. We applied this approach to a set of HSV-1 clinical swabs and paired single-passage culture isolates and saw limited sequence evolution with 14 out of 17 specimens being completely identical in the UL-US regions. With a separate set of clinical samples, we compared HSV-2 sequences from swab-derived specimens sequenced after different methods of storage (swab in viral transport media or PCR buffer, single passage culture, extracted DNA) and again saw minimal sequence evolution across different specimen types. Together, these results show that low-passage clinical isolates are reflective of the viral sequences present in the lesion and can be used for phylogenetic analyses. We have also used this method to detect superinfection by unrelated HSV strains in single and temporal samples, illustrating the power of direct-from-specimen sequencing of HSV.
**Background** Group B Streptococcus is a leading cause of neonatal invasive disease, however, there is limited information on the invasive disease genotypes from Africa. This study aimed to investigate genotype diversity and antimicrobial resistance genes associated with among invasive GBS isolates collected over a 12 years period in Johannesburg, South Africa.  

**Methods** Whole genome sequencing was performed using illumina bio-sequencer and Nextera DNA kit. Whole genome multi-locus typing was used to determine the genetic diversity of Group B Streptococcus isolates. The presence of resistance genes and pilus islands were identified using PubMLST.  

**Results** Among 293 isolates, 17 genotypes were found with ST17 (36.51%) and ST23 (19.79%) as being dominant genotypes, followed by ST109 (16.3%), ST1 (4.09%), ST28 (4.09%) and ST10 (3.41%). The invasive disease isolates were mostly associated (90%) with cps-Y, -L, and -F capsular biosynthesis genes. Pilus islands (PI) identified included PI-2b (24.5%), PI-2a (26.2%), and PI-1 (28.6%) and in combination PI-1+2a (9.5%) and PI-1+2b (27.9%). Nearly 92.8% of invasive isolates had a Tet-M gene and 5.11% had erm gene in their genome.  

**Conclusion** The dominant sequence type of invasive disease isolates were ST17 and ST23. The presence of tetracycline resistance gene and PI-1 were observed among the most dominant genotypes of Group B Streptococcus.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 138

**Title:** Elucidation of Major Environmental Factors That Govern the Microbial Community Structure and Function in Acid Mine Drainage of Malanjkhand Copper Project, India

**Author:** A. Gupta, A. Dutta, J. Sarkar, M. K. Panigrahi, P. Sar;

**Block:** Indian Institute of Technology Kharagpur, Kharagpur, INDIA.

**Abstract:**

**Background:** Biological oxidation of the sulfidic ores in the mining region generates highly acidic mine drainage which is threat to an ecosystem as it contains high concentration of heavy metals and sulfate, hence considered to be an extreme environment for life. The present study is designed to understand the role of major environmental factors (pH, SO$_4^{2-}$ and DOC) in assemblage of microbial community structure and function. **Methods:** Acid mine drainage (water and sediment) samples collected from the Malanjkhand copper project, India were used in the present study. To attain this objective, geochemistry of the samples was thoroughly assessed followed by 16S rRNA based targeted sequencing and shot gun metagenome approach to understand the microbial diversity and function as well as statistical analysis was used to comprehend the role of environmental factors in shaping the microbial community composition. **Results:** The samples were found to be distinct in its geochemical parameters and were partitioned in to two pH regimes [low (1.9 < pH < 4.0) and high 4.0 < pH < 6.0)]. The low pH samples contained high sulfate and heavy metals concentration as compared to high pH samples. The microbial diversity of the low pH samples were dominated with highly acidic, Fe/S oxidizing taxa responsible for AMD generation whereas the high pH samples constituted of moderately acidophilic/neutrophilic microbial groups involved in diverse biogeochemical cycling and contained groups which could be used as a potent members for AMD attenuation. Canonical correspondence analysis established the role of pH, Fe, SO$_4^{2-}$, DOC etc. in shaping the structure of microbial community composition. Spearman correlation of the OTUs with pH, Fe, SO$_4^{2-}$ revealed that highly acidic and Fe/S oxidizing groups were found to be positively correlated with these parameters. To understand the metabolic potential of the microbial community, one sample from each pH regime was considered for shot gun metagenome based approach and results revealed that genes involved in diverse biogeochemical cycling were detected in both the samples. The high abundance of genes involved in S oxidation and pH stress were detected in low pH sample whereas sulfate reducing gene were found to be more in high pH sample. The genes involved in C fixation, nitrogen metabolism and heavy metal stress were detected in both the samples hence confirmed the function of these organisms under low organic carbon and heavy metal stress. **Conclusion:** The present study provides a deeper insight into the role of environmental variables in shaping the microbial community structure and function in acid mine drainage.
Title: Rapidly Accumulated Tobramycin Resistance by *Pseudomonas aeruginosa* in CF-like Acidic pH Environment

Author Block: Q. Lin, Y. Di; University of Pittsburgh, Pittsburgh, PA.

**Background:** Cystic fibrosis (CF) is a genetic disease with a loss of cystic fibrosis transmembrane conductance regulator (CFTR) function that leads to impaired airway host defense. Chronic infection and colonization by gram-negative *Pseudomonas aeruginosa* (*P. aeruginosa*), an opportunistic pathogen, contribute to high mortality rates in CF. While the airway surface liquid of CF patients becoming more acidic with aging, the prevalence of *P. aeruginosa* lung infection also gradually increases over time in CF patients from age 2 to 45 and *P. aeruginosa* eventually becomes the dominant bacterial strain colonized in the lungs of CF suffers. We previously demonstrated that the acidic CF lung microenvironment promotes *P. aeruginosa* biofilm formation and multi-drug resistance. But the effects of acidic CF lung microenvironment on tobramycin treatment-associated antibiotic resistance (AR) remains unknown. In this study, we hypothesize that the acidic microenvironment promotes faster and stronger tobramycin resistance compared to physiologically neutral pH non-CF lung microenvironment.

**Methods:** Planktonic and bead-transfer biofilm models were used for *P. aeruginosa* PA14 evolution study in pH 6.5 and 7.5 with or without tobramycin treatment. Bacterial whole genome sequence data were acquired by Next-Generation Sequencing (NGS) technology.

**Results:** Our results indicated that PA14 exhibited a rapid morphological change under acidic pH conditions. Acidic environment also stimulated faster and stronger PA14 tobramycin resistance compared to neutral pH conditions. NGS results showed that acidic environments elicited several DNA mutations that were likely pH-dependent.

**Conclusions:** Our results indicated that PA14 generated AR quickly under tobramycin treatment and the acidic lung microenvironment promoted even faster tobramycin resistance in the biofilm mold of growth. The pH-dependent DNA mutations are potential targets for future treatment in CF patients to effectively eliminate *P. aeruginosa* infection.
Title: Genomic Characterization and Phylogenetic Analysis of *Salmonella* Javiana Clinical Strains from Tennessee, 2017-2018

**L. K. Hudson**\(^1\), C. Moore\(^2\), L. Constantine-Renna\(^3\), X. Qian\(^2\), L. S. Thomas\(^2\), K. N. Garman\(^3\), J. R. Dunn\(^3\), T. G. Denes\(^1\);

\(^1\)Department of Food Science, University of Tennessee, Knoxville, TN, \(^2\)Tennessee Department of Health, Division of Laboratory Services, Nashville, TN, \(^3\)Tennessee Department of Health, Nashville, TN.

**Abstract:**

**Background:** *Salmonella* Javiana is the fourth most common serovar of *Salmonella* found to cause illnesses in Tennessee (TN), but is geographically clustered in the western region. Almost two-thirds (63%) of *S*. Javiana clinical isolates from January 2017 through June 2018 were from counties in west TN. The objectives of this study were to retrospectively examine the genomic population structure of Javiana isolates from patients in TN in 2017-18 and describe epidemiological features among clades of case-patients identified. **Methods:** Biosample numbers and metadata for *S*. Javiana (n=61) clinical isolates from TN collected January 2017 to June 2018 were provided by the Tennessee Department of Health. Raw reads were downloaded from the NCBI SRA database, trimmed using Trimmomatic, and quality checked using FastQC. An appropriate reference assembly was chosen (BioSample SAMN01832085) and downloaded from the NCBI refseq database. hqSNPs were identified using the CFSAN SNP pipeline and the resulting matrix was used to construct a neighbor-joining tree with Mega7. Additionally, trimmed reads were assembled using SPAdes and contigs annotated with Prokka. Assembly statistics were generated by BBMap, SAMtools, and QUAST. SeqSero was used to confirm serotype designations. **Results:** Two distinct major clades of interest were identified (clades 1 and 4), each with geographical clustering, along with 2 minor clades. Major clades were defined as containing five or more isolates and minor clades as containing less than five. Clades may represent or contain epidemiological clusters. Clade 1 consisted of 23 isolates, with almost all (n=21) being isolated from patients in the western region of TN and the majority isolated in a single county (Shelby; n=13). Isolates from Clade 1 were collected over a period of approximately nine months. Clade 4 consisted of 20 isolates, mostly isolated from counties in west TN (n=13), that were collected over a period of approximately one year. There is a notable subclade of seven isolates within clade 4 (subclade 4A), all from four rural counties in one geographic location (Carroll, Gibson, Crockett, and Madison counties). These were collected over a period of about five months and the majority of isolates from this subclade were collected from patients that were male (86%) and adults (86%). **Conclusions:** The clustering patterns of *S*. Javiana isolates with a large portion originating in western TN, together with the timeline of the isolation dates and SNP differences, may indicate that many of these are environmentally acquired. Further investigation of epidemiological data and possible environmental sources may identify the source of illness and possible preventive strategies. In addition, information gained about the population structure of this serovar provides guidance for selecting SNP distance thresholds used to identify clusters that may be of epidemiological significance.
Identifying Metabolically Active Bacteria in Tobacco Products with DNA Labeling and Next-generation Sequencing

S. Chattopadhyay\(^1\), L. Malayil\(^1\), E. Mongodin\(^2\), A. Sapkota\(^1\);
\(^1\)University of Maryland, College Park, MD, \(^2\)University of Maryland, Baltimore, MD.

The advent of the Family Smoking Prevention and Tobacco Control Act, implemented by the U.S. Food and Drug Administration, has resulted in the need to improve our understanding of the microbial constituents of tobacco products. 16S rRNA gene sequencing techniques have enabled us to gain insights into identifying non-culturable bacteria present in tobacco products. These sequencing techniques generate massive amounts of unbiased data but are unable to determine what proportion of identified bacterial communities are live and active. To bridge this knowledge gap, our study aimed to identify and quantify the metabolically active bacterial communities in commercially-available tobacco products. We tested 14 tobacco products: 4 brands of cigarettes, 4 brands of little cigars and 6 brands of hookah, each with three distinct flavors. For each product, 0.2g of tobacco was treated with either i) Propidium monoazide (PMA), allowing detection of viable bacteria by inactivating DNA that is not contained within an intact cell membrane, or ii) 5-bromo-2'-deoxyuridine (BrdU), allowing detection of proliferating cells, or left untreated (control samples), after which total genomic DNA was extracted. BrdU samples were immunocaptured and all samples underwent PCR of the 16S rRNA gene followed by sequencing on Illumina HiSeq. Downstream analyses were performed using QIIME and R. Overall, 1,242 species-level operational taxonomic units (OTUs) were identified from more than 11 million sequences across 88 samples. Alpha diversity analysis (Observed and Shannon indices) showed significant (p<0.005) differences between bacterial communities among BrdU-treated, PMA-treated and control samples across all tobacco products. In addition, flavoring of tobacco products also showed significant effects on bacterial community composition. Beta diversity analysis comparing products using Bray-Curtis dissimilarity also identified significant differences between BrdU- and PMA-treated samples (ANOSIM R= 24.4%, p<0.001). Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria were the top phyla identified across all products. Our data confirms that tobacco bacterial communities are diverse and differ across brands and products. This study is the first to characterize the presence of a metabolically active fraction of bacterial communities residing within these products, which are affected by tobacco flavors.
**Background.** Shiga toxin-producing *Escherichia coli* (STEC) is a leading foodborne pathogen with a diverse genetic background that contributes to variation in disease presentation and severity. The use of WGS allows for a more comprehensive genomic profiles to be rapidly and easily obtained for comparing between isolates and for identification of genetic factors associated with disease outcomes. 

**Methods.** STEC isolates were obtained from 2010-2014 as part of an active surveillance system, which included four hospitals in Michigan. Wizard Genomic DNA purification and Illumina Nextera XT kits were used followed by sequencing using the Illumina MiSeq platform. *De novo* genome assembly was performed with Spades following trimming and quality checking with Trimmomatic and FastQC. **Results.** WGS data is available for 477 STEC isolates (33 O157 and 444 non-O157) from Michigan. A workflow was developed with bioinformatic scripts to extract the molecular serotype, virulence and resistance gene profiles, and multilocus sequence type (ST) for each strain. SNPs specific for one of nine clades were extracted from the 33 O157 strains as were CRISPR spacer regions to demonstrate a high level of diversity with multiple unique gene profiles. 45 non-O157 serogroups were identified and the isolates were grouped into 54 STs with 4 new STs identified. The O157 strains grouped into 5 clades, with the majority (n=14) belonging to clade 8. **Conclusion.** Use of WGS to characterize 477 STEC isolates has demonstrated that strains recovered from patients in Michigan are diverse and that specific gene profiles can be associated with epidemiological data. Comparative genomic analyses of STEC and other foodborne pathogens are important to identify key profiles that are most important for severe infections, and to validate existing subtyping methods.
Title: Genome Sequences of *Bacillus sporothermodurans* Strains Isolated from Ultra High Temperature (UHT) Milk

**Author Block:**
R. Owusu-Darko¹, M. Allam², S. D. Oliveira³, C. A. Ferreira³, S. Grover⁴, S. Mtshali², A. Ismail², E. M. Buys¹;

¹University of Pretoria, Pretoria, SOUTH AFRICA, ²National Institute for Communicable Diseases, Johannesburg, SOUTH AFRICA, ³School of Sciences, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, BRAZIL, ⁴Dairy Microbiology Division, Molecular Biology Unit, National Dairy Research Institute, Karnal, INDIA.

*Bacillus sporothermodurans*, first isolated in ultra-high temperature (UHT) milk, is a thermo-resistant, Gram-positive bacterium that can produce highly heat resistant endospores (HRS), that may survive UHT heat treatments. We sequenced four genomes of *B. sporothermodurans*, including for the first time, both heat resistant and non-heat resistant strains. The size of the genomes ranges from 3.4 Mb to 3.9 Mb with an average G + C content of 36 % and the number of coding sequences ranging from 3768 to 4558. Our research also shows that both heat resistant and non-heat resistant strains have similar compliment of heat resistance genes, the hrcA-dnaK-dnaJ-grpE operon and biofilm formation of the TasA and homologs. The whole genome sequence of three of the four sequenced *B. sporothermodurans* strains have the *Listeria sp.* pathogenicity island LIPI-1, presumably obtained through horizontal gene transfer. Evolutionary trends of *B. sporothermodurans* suggest a common ancestor originating from the gut of insects or Arachnids like its closest phylogenetic neighbor, *Bacillus oleronius*. The draft genomes carried out on the Illumina MiSeq system will enhance our understanding of the genes and pathways responsible for heat resistance and biofilm formation which is of prime importance to the dairy industry. It will also allow for pangenome studies which are ongoing and the evolutionary relationships with other *Bacillus* species of concern to the food industry. PacBio sequencing to start in earnest will allow to fill out the gaps in the genomes undertaken through the MiSeq platform.
Session: **Poster Session B**  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 145  
**Title:** Dissection of The Mobilome of Carbapenem-Resistant *Klebsiella pneumoniae* (CR-Kpn) Using Short and Long Read Assemblies: A Prospective Study in Houston, TX  
**Author Block:** W. C. Shropshire, A. Q. Dinh, H. Ecklund, A. Wanger, W. Miller, D. Panesso, T. T. Tran, C. A. Arias, B. M. Hanson; UTHealth, Houston, TX.  

**Abstract:** *Klebsiella pneumoniae* (Kpn) is a gram-negative pathogen that is responsible for nosocomial infections leading to significant morbidity and mortality worldwide. Mobile genetic elements (MGEs; e.g. plasmids and transposons) are of crucial importance for these organisms to adapt and evolve. Exchanging of accessory genes that confer resistance to antimicrobials is particularly important in clinical settings. The application of short- and long-read sequencing platforms with next generation sequencing (NGS) bioinformatic tools permits high resolution of these complex resistance elements which otherwise prove difficult using one sequencing platform alone. Here, we describe the genomic profiles of 95 CR-Kpns belonging to clonal group (CG) 258 and non-CG258 collected in hospitals across Houston, TX from May to December 2017.  

**Methods:** Libraries were prepared with Nextera XT DNA Library Prep Kit (Illumina) and Rapid Sequencing Kit (SQK-RBK004, Oxford Nanopore Technologies, ONT). Sequencing platforms used were the MiSeq and HiSeq 4000 (Illumina) and GridION X5 (ONT). A custom pipeline was developed for high-throughput data QC, processing, assembly, and analysis of Illumina data. Oxford Nanopore sequencing data was assembled using Canu v1.7.1 and polished with Illumina sequencing data using Pilon 1.22. Chromosomes and plasmids were circularized using Circlator 1.5.5.  

**Results:** Phylogenetic and multi-locus sequence typing (MLST) analysis on 95 Kpn samples revealed two predominant sequence types (STs) with 38/95 (40%) ST258 and 35/95 (37%) ST307. Short-read alignment analysis indicated that all ST307s had the extended spectrum beta lactamase (ESBL) *CTX-M-15* gene whereas it was only present in one ST258 isolate. Two non-ST258/ST307 isolates carried the genes encoding NDM-1 carbapenemase. Two representative isolates from each dominant clade and four non-ST258/ST307 isolates were sequenced using the GridION X5 platform with their plasmid and MGE structures closed and resolved respectively. Fox example, we were able to resolve transposon Tn4401a linked with *blaKPC-3* carriage within an ST307 isolate on a single ONT read. This allowed us to identify multiple isoforms and SNPs of genes initially identified through the abricate tool, i.e. CARD and PlasmidFinder, and gain greater detail of the MGE structures that carried them.  

**Conclusions:** Initial phylogenetic analysis revealed two clades, ST258 and ST307, which appear to dominate the multifocal prevalence within our Houston hospital setting. The application of two NGS sequencing platforms along with our custom bioinformatics pipeline allowed a complete elucidation of the MGE structures of interest as well as the resistance determinants of which these MGEs carried.
Session: **Poster Session B**

**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm

**Poster Board #:** 146

**Title:** Hybrid Sequencing and Assembly Reveals Genomic Diversity of Methicillin-susceptible *Staphylococcus aureus* (MSSA) from a Neonatal Intensive Care Unit (NICU) Surveillance Effort

**Author Block:** M. K. Annavajhala, W. Geng, A. C. Hill-Ricciuti, S. Ferguson, S. L. Stump, M. J. Giddins, M. Messina, P. Zachariah, D. A. Green, S. Whittier, L. Saiman, A. Uhlemann; Columbia University Medical Center, New York, NY.

**Background:** MSSA is a more prevalent NICU pathogen than methicillin-resistant *S. aureus* (MRSA), yet optimal MSSA infection prevention and control strategies are unclear. Given the ubiquity of MSSA as a human colonizer, neonatal acquisition likely occurs through multiple routes during delivery and close contact with parents and healthcare providers. However, the introduction and potential local spread of MSSA and the role of systematic decolonization for infants colonized with MSSA remain incompletely understood. Here, we used short- and long-read whole-genome sequencing (WGS) to define the diversity of MSSA during an ongoing NICU surveillance effort and aimed to identify genomic features potentiating the spread of prevalent MSSA clones.

**Methods:** Infants hospitalized in a 75-bed university-affiliated level III-IV NICU over an 18-month period were screened twice monthly for MSSA-positive clinical and/or pooled four-site surveillance cultures. We typed isolates using PCR and sequenced the most prevalent *spa* types using Illumina WGS (n=107). We used SRST2 for in silico multilocus sequence typing and antibiotic resistance gene and plasmid replicon typing. Oxford nanopore sequencing and hybrid assembly for each MLST type was used to create optimal reference genomes. We included previously published data in phylogenetic analyses of core genome SNPs to identify the evolutionary history of major clones.

**Results:** We collected 466 MSSA isolates from 297 infants. MSSA *spa* types identified (80 in total) included t279 (n=86), t1451 (n=21), and t571 (n=10). Of note, t1451 and t571 belong to ST398, a common clindamycin-resistant MSSA in our local community. In contrast, t279 (CC15) has not been encountered in community surveillance efforts yet increased in the NICU during the study period. We used nanopore sequencing of the oldest 2016 CC15 isolate to generate a reference genome. Compared to publicly available genomes, our reference greatly reduced pairwise SNP distances and allowed for more accurate phylogenetic inferences. ST398 NICU isolates formed three clusters with closely related community isolates, suggesting community members and NICU reservoirs as sources of acquisition in neonates. CC15 comprised two clades of closely related isolates (&lt; 100 SNPs) distinct from known community MSSA, pointing to clonal expansion within the NICU. Almost all CC15 also harbored *mupA*-encoding plasmids, including the reference isolate, indicating potential proliferation due to decolonization efforts with mupirocin.

**Conclusions:** MSSA in our NICU exhibited substantial genetic heterogeneity. Comparative genomics indicate genotype-specific pathways of introduction and spread of MSSA, including potential community- (ST398) or healthcare- (CC15) associated sources. Antibiotic resistance may play an important role in dissemination of CC15. Future surveillance efforts could benefit from routine genotyping.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 147
Title: Microfluidic NGS Sample Preparation for High-Throughput Epidemiology
Author: S. Kim, G. Lagoudas, P. Blainey;
Block: The Broad Institute of MIT and Harvard, Cambridge, MA.

Abstract: While low-cost DNA sequencing is transforming biological research and discovery, preparing large sample sets for sequencing with minuscule starting material is now the limiting factor in many applications. Here we introduce a polydimethlysiloxane (PDMS) microfluidic device that automates the key steps in whole genome NGS sequencing sample preparation, integrating lysis, fragmentation, adapter tagging, purification, and size selection of 96 samples in parallel. We applied our device to process about 5000 whole genome sequencing libraries of Pseudomonas aeruginosa clinical isolates, methicillin resistant Staphylococcus aureus, Mycobacterium tuberculosis, soil microbes, and human gut microbes using dramatically reduced sample input and reagent quantities. These microfluidic libraries showed excellent coverage for variant calling, phylotyping, metabolic profiling, and metagenomic analysis performance from only 10,000 cells (50 picograms of genomic DNA). Our method will enable high-throughput processing of samples for shotgun sequencing with broad application to basic science and clinical medicine.
Session: **Poster Session B**
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 148
Title: PathOGiST: Calibrated Multi-criterion Genomic Analysis for Public Health
Author Block: P. Feijao¹, M. Katebi², E. Lasalle², H. Yao³, S. La¹, M. Nguyen¹, C. Chauve¹, L. Chindelevich¹; ¹Simón Fraser University, Vancouver, BC, CANADA, ²École Polytechnique, Paris, FRANCE, ³École Polytechnique, Paris, FRANCE.

As public health organizations start to rely on whole-genome sequencing (WGS) data for infectious disease surveillance and outbreak investigations, two main issues emerge from the use of WGS data for genotyping. First, methods for differentiating outbreak-related strains from sporadic strains are often based on a single type of genomic variation. This approach captures only a limited amount of the genomic variability and tells only a partial story of the organism’s evolutionary history. Second, WGS-based sample clustering algorithms are often not calibrated, meaning that the determination of clustering thresholds or subtyping cutoffs is still mostly arbitrary. There are many forces driving pathogen evolution and as a result, using the wrong set of variants or the wrong cutoffs may mislead the investigation of a pathogen outbreak. We address this issue by developing PathOGiST, that implements and integrates existing and novel genomic variant calling algorithms from WGS data (SNPs, Multi locus sequence typing and copy number variations), together with clustering algorithms based on a multi-criterion genome dissimilarity measure using various kinds of genomic variants. Final steps include the calibration of the statistical models and algorithms using large reference sets of selected pathogen genomes from epidemiologically confirmed outbreak strains. The PathOGiST pipeline will be implemented both as a standalone and as a Galaxy tool, and will be part of the IRIDA platform, making it available to public health workers in Canada and around the world.
**Session:** Poster Session B  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 149  
**Title:** Worflows for detection in methylomes in ONT and Pacbio  
**Author Block:** L. A. Arteaga-Figueroa, V. Villegas-Escobar, J. Correa-Alvarez; Universidad EAFIT, Medellin, COLOMBIA.

**Abstract:**

**Background:** The Third generation sequencing (TGS) technologies present many advantages in comparison with Next generation sequencing (NGS), mainly because its capability to produce long reads, detect base modifications, RNA sequencing, superior performance on repeated regions. Many comparisons are found regarding the performance, but few compare the capability to detect DNA modifications, available software performance, and final annotated methylome. In this study, we provide workflows for the detection of 6mA and 5mC in Nanopore (ONT) and SMRT sequencing data, and a comprehensive comparison between them. **Methods:** Data obtained with ONT and PacBio from Bacillus subtilis project (EACB0015) were used to design workflows for the base detection analysis (considering the most tools available). These workflows compare intensively mappers and modified base detectors. For 6mA, we implemented mCaller, Tombo, and ipdSummary (kineticTools); for 5mC, we implemented mCaller, Tombo, Nanopolish and ipdSummary. Homescrpts in python were used for output analysis and graphs. **Results:** In terms of mapping, for PacBio, so far, only Blasr was able to output the .bam necessary for the analysis. For ONT data, Graphmap and Minimap2, although Graphmap was more accurate than Minimap2 for high error datasets, Minimap2 performed better; we are currently testing lordFAST. For 5mC detection in ONT data, we found that Tombo is the most sensitive, nearly followed by Nanopolish. For 6mA, Tombo was also the most sensitive. ipdSummary failed many times to assign identity to modified bases, and outputted many less bases than ONT software for both 6mA and 5mC. Additionally, detection of modified bases in the same position (in dimers like TA, GC) but in different strands were also observed. The missidentification of 4mC remains as an issue through the analysis, the later chemistry of PacBio does not have an appropriate software for 4mC detection. When we compared ONT for 4mC with 5mC ONT detections, some positions matched, and as aforementioned, a big part of the output of ipdSummary does not assign identity to all the modified bases, we are still working in the identification of 8oxoA, 8oxoG, and other kinds of mC with the purpose of identify possible false positives for 5mC in ONT results. **Conclusions:** Oxford Nanopore sequencing showed to be more sensitivethat PacBio sequencing for 5mC and 6mA detection accordingto our results.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 150
Title: Metagenomic MinION and Illumina Sequencing for Surveillance of Cholera and Other Waterborne Pathogens in Haiti
Author: B. Stebbins, S. Hung, C. Martin, T. Ford; UMass Amherst, Amherst, MA.
Block: The recent deadly cholera outbreak after the 2010 earthquake in Haiti prompted the need for a portable system that will allow for rapid pathogen identification without requiring expensive laboratory resources. To address the need, we are testing the field suitability of the hand-held sequencer called the MinION (Oxford Nanopore Technologies) for the metagenomic assessment and detection of waterborne pathogens such as *Vibrio cholerae*. In the initial testing phase, we collected and tested water samples for coliforms from the Mill and Fort Rivers in Amherst and Hadley, Massachusetts. DNA was then isolated, spiked with *V. cholerae* DNA, and prepared via the DNA ligation method for sequencing. Initial MinION sequencing followed by bioinformatic analysis of the data using WIMP, OneCodex, and CosmosID (Rockville, MD) platform detected *V. cholerae* spike-in but not the coliforms unless the samples were enriched for these species. Further optimization of the portable system is being undertaken to allow for future field metagenomic applicability. We will also be using Illumina MiSeq sequencing for comparison of read accuracy.
Abstract:
Session: **Poster Session B**

**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm

**Poster Board #:** 151

**Title:** HAIviz: Visualizing Genomic Epidemiology Data of Healthcare Associated Infections

**Author Block:** B. Permana\(^1\), B. M. Forde\(^1\), L. Roberts\(^1\), P. Harris\(^2\), S. A. Beatson\(^1\); \(^1\)University of Queensland, St Lucia, AUSTRALIA, \(^2\)UQ Centre for Clinical Research, Brisbane City, AUSTRALIA.

**Abstract:**

Visualization is an essential aid for communicating genomic epidemiology of infectious disease that often employs complex epidemiological and genomic information. Over the past few years, several tools such as Microreact, PhyloGeoTool, and Nextstrain have been developed and demonstrated the benefit of data integration and interactive visualization. However, these tools mainly focus on global epidemiology and phylogeography. Applications that feature specific data related to Healthcare Associated Infection (HAI) such as the patient's bed movements, hospital room layout, and infection transmission network remain unavailable.

Here we present HAIviz, an interactive single web page application for visualizing genomic epidemiology of HAI. It was developed using popular web technologies to allow infection control professionals, epidemiologist, clinicians, and public health decision-makers explore potential insights from Whole Genome Sequencing (WGS) data and epidemiological information. HAIviz allows users to display a detailed hospital map, integrated with patient metadata, phylogenetic trees and transmission networks. Users can create single or multiple visualization windows by uploading their own dataset in the required format; metadata and transmission files in Comma Separated Value (CSV), maps in GeoJSON, and trees in Newick. Each generated window is independently arranged, giving users the freedom to display their preferred information.

HAIviz is freely available at the URL http://haiviz.beatsonlab.com and can be accessed using any modern browser. As a client-side application, HAIviz perform all computational process in the user's machine with no information posted to the server, making it inherently private, secure, and scalable. Currently, HAIviz is accessible as a standalone visualization application that works with the input files created by a separate workflow. In the future, HAIviz aimed to be an integrated system with WGS-based epidemiology and bioinformatics pipeline, promoting a real-time HAI surveillance and investigation framework.
**Abstract:**

**Background:** Countries are experiencing a serious public health threat and major obstacle to disease control due to excess antibiotic use and drug-resistant *Mycobacterium tuberculosis*. Single nucleotide polymorphisms (SNPs) and groups of virulence genes will be used for genotyping. Single nucleotide polymorphism could be the most valid markers due to the very low level of homoplasy and that they are ideally suited for defining phylogenetic grouping with very high confidence. **Objectives:** To analyse approximately 800 *M. tuberculosis* whole genome sequencing (WGS) data deposited in a web-based comprehensive information system PATRIC website and select drug resistance and susceptible genomes for genotyping. To create database of virulence gene mutation catalogues based on *M. tuberculosis* genome deposited from Brazil, China and South Africa. **Research Methods:** In this work a bioinformatics analysis of virulence genes from 303 whole genome sequencing of *M. tuberculosis* was performed from Brazil (n=2), China (n=23), India (n=238), Russia (n=259) and South Africa (n=278) downloaded from PATRIC database and analysed on CLC Genomics Workbench 11. **Results:** A bioinformatics analysis of *M. tuberculosis* WGS showed that out of 15 tested genes (*mazF3; vapB17; vapC47; higA; vapC37; vapC38; vapC6; mazF8; vapC3; mce3B; cyp125; vapC25; vapB34; mce3F; vapC46*) only from gene *vapC3* did not have mutations. Several genes were found to carry SNPs that correlate with specific genotypes. Using *vapC37* and *vapC38* we observed Beijing lineage and its sublineages which are associated with drug resistance and elevated virulence. Mutations obtained were specific for lineage and sublineage level. *VapC3; vapC38; and mazF8* genes were associated with LAM1, 2,9 and LAM4/F15/KZN sublineages. **Conclusions:** The constructed SNP reflected the evolutionary relationship between lineages. In future we will need to establish a South African *M. tuberculosis* catalogue of SNPs in virulence gene specific to the F15/LAM4/KZN lineage this will complement the diagnostic pipeline using WGS data for drug resistance detection and lineage determination. We will determine a list of conserved genes that can be used for future development of DNA vaccines.
**Background:** As the cost of sequencing has declined, clinical diagnostics based on next generation sequencing (NGS) have become reality. Diagnostics based on sequencing will require rapid and precise mapping against redundant databases because some of the most important determinants, such as antimicrobial resistance and core genome multilocus sequence typing (MLST) alleles, are highly similar to one another. In order to facilitate this, a novel mapping method, KMA (k-mer alignment), was designed. KMA is able to map raw reads directly against redundant databases, it also scales well for large redundant databases. KMA uses k-mer seeding to speed up mapping and the Needleman-Wunsch algorithm to accurately align extensions from k-mer seeds. Multi-mapping reads are resolved using a novel sorting scheme (ConClave scheme), ensuring an accurate selection of templates.

**Results:** The functionality of KMA was compared with SRST2, MGmapper, BWA-MEM, Bowtie2, Minimap2 and Salmon, using both simulated data and a dataset of *Escherichia coli* mapped against resistance genes and core genome MLST alleles. KMA outperforms current methods with respect to both accuracy and speed, while using a comparable amount of memory.

**Conclusion:** With KMA, it was possible to map raw reads directly against redundant databases with high accuracy, speed and memory efficiency. **Availability:** KMA is implemented in C, and is freely available at: https://bitbucket.org/genomicepidemiology/kma and as web-service at: https://cge.cbs.dtu.dk/services/KMA/.
Population Genomics Identified *Salmonella* Newport ST45 as the Main Driver of Emerging Multidrug Resistance

**Author** M. Yue¹, S. Rankin², D. Schifferli², W. Fang¹, H. Pan¹;

¹Zhejiang University, Hangzhou, CHINA, ²University of Pennsylvania, Philadelphia, PA.

*Salmonella* is one of the most important foodborne pathogens in the world, the emerging of multidrug-resistant *Salmonella* clones pose a significant threat for veterinary public health and food safety. However, the genetic and/or evolutionary pressure for the selection of antibiotic-resistant pathogens in food animals remains poorly understood. The aim of this study was a global investigation of the population diversity of *S.* Newport isolates by studying the MLST of 2,250 isolates. Three clades were identified that correlated with the niches/origins of isolation (human, animal, and environment). Sequence analysis of 1,855 *S.* Newport genomes identified Sequence Type 45 (ST45) as the predominant clone among the animal isolates (87%), but only in 9% of the isolates from human infections. ST45 isolates carried multiple plasmids, the majority (> 90%) had a unique IncA/C plasmid that ranged in size from 80 to 200 kb. The plasmid carried genes responsible for multidrug resistance, including *floR, tetAR, strAB, sul, mer*, and *bla*. Importantly, three Chinese strains carried the *mcr-1* gene, that confers plasmid-mediated resistance to colistin, one of a number of last-resort antibiotics for treating Gram-negative bacterial infections. A genome-wide association study (GWAS) correlated chromosome regions or genetic variations with maintenance of an IncA/C plasmid in ST45 isolates. An additional investigation of the minimum inhibitory concentration (MIC) of 27 antibiotics in 3,728 isolates isolated from the food-chain (food-animals, retail meats, and humans) suggested that AR *S.* Newport from humans have multiple, but distinct origins. Animal and retail-meat isolates are distinct from > 92% of the human isolates by their antibiotic-resistance patterns. Taken together, our findings suggest *S.* Newport ST45 is the dominant clone in food-animals in the world. The GWAS data will serve to investigate genetic determinants that contribute to the maintenance of this clone in food-animals.
BacPipe: A Rapid, User-Friendly Whole Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology and Outbreak Detection

B. Xavier, M. Mysara, M. Bolzan, C. Lammens, S. Kumar-Singh, H. Goossens, S. Malhotra-Kumar;
University of Antwerp, Wilrijk, BELGIUM.

Despite rapid advances in whole genome sequencing (WGS) technologies, their integration into routine microbiological diagnostics and infection control has been hampered by the need for downstream bioinformatics analyses that require considerable expertise. We have developed a comprehensive, rapid, and computationally low-resource bioinformatics pipeline (BacPipe) that enables direct analyses of bacterial whole-genome sequences (raw reads, contigs or scaffolds) obtained from second and third-generation sequencing technologies. BacPipe is an ensemble of state-of-the-art, open-access tools for quality verification, genome assembly, annotation, and identification of the bacterial genotype (MLST, emm typing), resistance genes, plasmids, virulence genes, and single nucleotide polymorphisms (SNPs). The outbreak module (SNPs and patient metadata) can simultaneously analyse many strains to identify evolutionary relationships and transmission routes. Importantly, parallelization of tools in BacPipe considerably reduces the time-to-result. Validation of BacPipe using prior published WGS datasets from hospital, community and food-borne outbreaks and from transmission studies of important pathogens demonstrated the speed and simplicity of the pipeline that reconstructed the same analyses and conclusions within a few hours. We believe this fully automated pipeline will contribute to overcoming one of the primary hurdles to WGS data analysis and interpretation, facilitating its application for routine patient-care in hospitals and public-health and infection-control monitoring.
Rapid Extraction of Single-copy Core Genes for Species Delimitation

S. Wittouck¹, S. Wuyts¹, C. Meehan², V. van Noort³, S. Lebeer¹;
¹University of Antwerp, Antwerp, BELGIUM, ²Institute of Tropical Medicine Antwerp, Antwerp, BELGIUM, ³KULeuven, Leuven, BELGIUM.

Background: Many analyses in comparative genomics and phylogenetics rely on single-copy core genes (SCGs): genes present in all genomes of interest in exactly one copy. Current strategies to obtain SCGs are either slow or rely on pre-computed marker genes, either universal or lineage-specific.

Methods: We developed a tool for the rapid extraction of SCGs from large sets of genomes in linear time. The tool works by first identifying candidate SCGs on a random subset of “seed” genomes with OrthoFinder, which uses an approach based on all-vs-all blastp and MCL. This is followed by a search for those candidate SCGs in all genomes using HMMER. Finally, a score cutoff is determined per candidate SCG to optimize for single-copy presence and only candidate SCGs present in nearly all genomes are retained. We apply our tool to all 2110 publicly available genomes that belong to the Lactobacillus Genus Complex (LGC). We show the applicability of the obtained SCGs by using them for 1) quality control of all genomes, 2) species delimitation based on pairwise single-copy core nucleotide identities (SCNIs) and 3) phylogeny inference using one representative genome per species. In addition, we compare our SCNI-based species delimitation with ANI and TETRA based species delimitations.

Results: On a subset of 200 LGC genomes, we show that our tool identifies similar SCGs as full gene family clustering, but is faster. In the full dataset of 2,110 genomes, we identify 422 SCGs sensu lato. Using those, we find that 1,980 genomes are of high quality based on filters of < 5% missingness and < 5% contamination. The pairwise SCNI and ANI similarities are strongly correlated above and slightly below the species threshold, while, surprisingly, they are very weakly correlated for more distantly related genomes. Species delimitation of the high-quality genomes results in the identification of thirteen “new” species, in the sense that it is not yet known that genomes of those species are publicly available. Some of those genomes are annotated as other species but are too distant from their type strain to be classified as such, while others are annotated as unclassified on the species level. The phylogeny of the species shows that the new species are spread across the LGC tree, with some being closely related to known species, while others are more distant.

Conclusions: Our tool for rapid extraction of SCGs yields similar results as current methods and is much faster. We suggest the SCNI similarity as an alternative for ANI since it can be determined rapidly for large datasets, results in very similar species boundaries and might more accurately represent genome distances for more distantly related genomes. Finally, we identify thirteen new species among publicly available LGC genomes.
**Title:** Genomic Epidemiology of *Vibrio cholerae* O1 in Haiti: A Switch from the Ogawa to Inaba Serotype

**Author Block:**

*Vibrio cholerae* is the causative agent of the disease cholera. This bacterium is ubiquitous in aquatic environments and toxigenic *V. cholerae* O1 may serve as a source for recurrent cholera epidemics around the globe. In January 2010, a massive earthquake struck Haiti, causing severe damage to the public health infrastructure. Then in October 2010, cholera appeared in Haiti for the first time in over 150 years. Previous studies show that the early cases of cholera in Haiti are consistent with a single-source introduction of *V. cholerae* O1 from Nepalese U.N. peacekeeping troops sent after the earthquake. After the initial epidemic waves, cholera may now be endemic in Haiti, showing seasonal outbreak patterns associated with the rainy season. This clonal, single-source introduction of *V. cholerae* O1 presents a unique opportunity to study the evolution and selective pressures acting on this microorganism. By performing phylodynamic analysis with genome-wide single nucleotide polymorphisms (SNP), we are able to investigate the ongoing cholera epidemic occurring in Haiti and the underlying evolutionary processes and selective pressures at a remarkable resolution. Since the start of the cholera outbreaks in 2010, the dominate serotype of *V. cholerae* O1 circulating in Haiti was the Ogawa serotype. Then in 2015, Inaba became the dominant serotype in Haiti. The main driver causing the switch from the Ogawa to the Inaba serotype is by a nucleotide substitution in the *wbeT* gene. Though the switch from the Ogawa to the Inaba serotype is a common phenomenon in the genome of *V. cholerae* O1, if the Ogawa serotype still remains dominate in the population and an outbreak of the Inaba serotype occurred, this could have been caused by a separate introduction into the population. Previous studies have shown that the Inaba serotype has been present in Haiti since 2012 but it has never propagated and become established as the dominate serotype circulating in Haiti. By using genome-wide SNPs to perform our analysis, we are able to assess potential evolutionary changes and selective pressures that are occurring in the *V. cholerae* O1 genome to generate this switch in serotype. Our results suggest that the *V. cholerae* O1 strains currently circulating in Haiti have evolved from their initial clonal, single-source outbreak of the Ogawa serotype to the new, unintroduced Inaba serotype.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 158
Title: LINbase: A Fast and Precise Whole Genome-Based Web Tool for Bacterial Pathogen Identification and Tracking
Author: L. Tian, L. S. Heath, B. A. Vinatzer; Virginia Tech, Blacksburg, VA.

Abstract: The current pragmatic approach to bacterial taxonomy provides clear classification guidelines to determine if a bacterial isolate belongs to an already named species. It also provides clear nomenclature rules on how to name new species. However, species descriptions do not reveal the extent of genetic and phenotypic diversity within species and current taxonomy does not provide any general guidelines or rules for intraspecific classification. This is highly problematic in the case of bacterial pathogens, including foodborne pathogens, since most pathogen species contain non-pathogenic strains and even pathogenic strains can be separated into different intraspecific groups based on genomic and phenotypic features. Whole genome sequencing (WGS) has shown considerable potential to facilitate the detection of foodborne disease outbreaks and origin tracking by increasing the discriminatory power compared to molecular methods such as pulsed-field gel electrophoresis (PFGE), multiple-locus variable number tandem repeat analysis (MLVA) and multi-locus sequence typing (MLST). Life Identification Numbers (LINs) have been shown to reflect phylogenetic relationships and to provide a system for classification at - and below - the species level. At the same time, LINs greatly improve identification of bacterial isolates because LINs can identify bacterial isolates as members of species or members of intraspecific groups or members of any other defined groups, e.g., isolates simply belonging to the same disease outbreak. LINbase is a Web tool that implements LINs for classification and precise identification of bacteria. In combination with fast algorithms and a user-friendly Web interface, LINbase will not only provide users with the ability to precisely identify any bacterial isolate based on its genome sequence within minutes, but also to determine outbreak-association.
Next Generation Sequencing to Investigate Nosocomial Transmission of Influenza

D. Frampton\textsuperscript{1}, R. Blackburn\textsuperscript{1}, C. Houlihan\textsuperscript{1}, C. Smith\textsuperscript{1}, Z. Kozlakidis\textsuperscript{1}, S. Hue\textsuperscript{2}, A. Hayward\textsuperscript{1}, E. Nastouli\textsuperscript{1};

\textsuperscript{1}UCL / Farr Institute for Health Informatics Research, London, UNITED KINGDOM, \textsuperscript{2}London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM.

**Overview:** Evidence-based infection control of nosocomial influenza has the potential to offer substantial human health improvements and financial cost-savings. However, few studies have examined nosocomial influenza transmission outside the narrow context of suspected outbreaks. This study is one of the first to apply full genome sequencing to examine influenza transmission in hospital settings and to compare genomic clusters defined at the level of the full genome with cases that were epidemiologically linked in time and space (hospital ward or clinic). Our findings exemplify the use of full genome sequencing for hospital surveillance of transmission, showing the technique can identify distinct transmission chains with substantially greater resolution than can be achieved through classical epidemiological investigation. We show that an important proportion of hospitalized influenza cases (at least one in eleven) lead to onward transmission usually involving short chains of transmission (average length of 3 cases). Many transmission chains cannot be explained by known contact between individuals suggesting “missing links” in the chain due to under-ascertainment of influenza cases in patients and a potential role for staff and/or visitors (who were not sampled) in transmission. **Methods:** All influenza samples from inpatients, outpatients and A&E attenders at a single hospital were included between September 2012 and March 2014. Clinical records were used to define patients with suspected nosocomial infection with possible transmission inferred from timing of first positive sample (relative to admission) and spatio-temporal links to other infected patients. Sequencing was by Illumina MiSeq. **Results:** 50 of 214 cases were part of genetically defined transmission chains amongst hospitalised patients. The proportion in genetic transmission chains was substantially higher for patients testing positive after 2 days of admission than those diagnosed soon after admission (p<0.001), and for those with spatio-temporal links compared to those without (p<0.001). The genetic distance between pairs of cases with spatio-temporal links was lower than that for pairs with no spatial links (p<0.001). Assuming each genetically identified cluster includes one community-acquired index case we estimate that 16% of hospital cases were due to nosocomial transmission. 1 in 11 cases seeded a new transmission chain comprising an average of 3 cases. **Conclusions:** Nosocomial influenza contributes significantly to hospital burden during outbreak seasons. Routine whole genome sequencing will support outbreak investigations and monitor the impact of infection and control measures.
Session: **Poster Session B**

**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm

**Poster Board #:** 160

**Title:** PiReT: Pipeline for Reference-based Transcriptomics

**Author Block:** M. Shakya, S. Feng, C. Lo, K. W. Davenport, B. Hu, P. S. Chain; Los Alamos National Laboratory, Los Alamos, NM.

**Abstract:**
Transcriptomics enables identifying genes and pathways that are differentially expressed in one condition over another, discovering small RNAs (sRNA), annotating transcribed genes, and characterizing alternative splicing. With the rapid advancement in sequencing technologies providing unprecedented throughput at an acceptable cost, many research laboratories have shown interests in applying transcriptomics for their research. However, most of the laboratories have found themselves continuously challenged by the lack of bioinformatics and statistical expertise needed to design, implement, and maintain computational workflows capable of analyzing transcriptomics data.

Here, we present PiReT, a pipeline for Reference-based Transcriptomics or PiReT, a one-of-a-kind reference-based transcriptomics solution that adopts an open architecture and is built upon web-based analysis platform of EDGE Bioinformatics to enable biologists with little or no computational knowledge to analyze their data. A typical transcriptomics workflow requires implementing an array of bioinformatics tools, each of which addresses a particular step in the analysis, e.g. quality control, alignment, fragment counting, statistical hypothesis testing, etc. PiReT effectively weaves together open source bioinformatics tools such as FaQCs, HISAT2, featureCounts, EdgeR, DeSeq2, etc. and presents it in an interactive web Graphical User Interface (GUI) where users can upload their raw data (fastq), customize steps of analysis, and produce biologist-friendly results (e.g. RPKM/FPKM/TPM, read counts, list of differentially expressed genes and pathway, etc.) and data visualizations within the GUI. It can perform metatranscriptome analysis like host and pathogens responses, detect sRNAs, and perform gene set enrichment or pathway analysis. PiReT can be used as a stand-alone workflow in command line and is also integrated into EDGE Bioinformatics.
Session: **Poster Session B**  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 161  
**Title:** Chronic Campylobacteriosis Outbreak Investigation in Great Apes Using Next-Generation Sequencing  
**Author Block:**  

D. Bandoy¹, E. Crook², N. Kong¹, C. Huang¹, B. Weimer¹;  
¹University of California, Davis, Davis, CA, ²Hogle Zoo, Salt Lake City, UT.  

**Abstract:** Background: Campylobacteriosis is one of the leading causes of diarrhea globally. While more than a thousand genomes have been published for *Campylobacter jejuni*, genomes from other *Campylobacter* species, like *C. hyointestinalis*, have been reported infrequently. To date only 18 *C. hyointestinalis* complete and draft genomes are available published, primarily from domestic and wild ruminants.  

**Methods:** Campylobacter was isolated and identified from a longitudinal surveillance strategy with feces using classical microbiological methods. Whole genome sequencing was done using the previously described protocol of the 100K Foodborne Pathogen Project using Illumina HiSeq X Ten instruments. Raw reads were assembled using CLC Genomics. Genome distance was computed using GGDC and the distance matrix values were used to generate a phylogenomic tree. Annotation was done using Prokka followed by pan-genome analysis using Roary that was visualized using Phandango. Whole genome sequences were analyzed using ABRicate and the online Comprehensive Antibiotic Resistance database (CARD) for antimicrobial resistance genes and virulence factors. Genomic islands were predicted using Island Viewer online tool and manual curation was performed using Mauve alignment.  

**Results:** In silico genome distance placed the isolates into distinct groups of host species of origin, indicating host species adaptation. This clustering corresponds to the host species specific set of genes as demonstrated by presence-absence variation (PAV) analysis using the pan-genome output. Only one isolate (BCW 9279) within the primate outbreak showed phylogenetic incongruence by clustering with the New Zealand deer isolates. The apparent phylogenetic incongruence has been determined to be due to horizontal gene transfer with genomic islands acquired from *Clostridioides difficile*. Further analysis revealed the existence of arsenical resistance genes which eventually was no longer identified in the post-antibiotic treatment genome sequences. This finding indicates a mechanism of genomic divergence with the acquisition of genomic islands that were negatively selected within the context of an ongoing outbreak and therapeutic intervention. Surprisingly, despite the great ape’s exposure to tetracycline treatment, *C. hyointestinalis ssp. hyointestinalis* resistome profiling revealed absence of any known antibiotic resistance genes. rRNA copy number comparison (one copy in all the primates versus three copies in the reference) suggests a possible mechanism of a carrier state with reduced metabolic activity.  

**Conclusion:** Whole genome sequencing revealed unprecedented resolution of longitudinal infection dynamics, revealing acute genomic island gain and loss due to negative selection pressure with antibiotic exposure. These findings highlight the genomic flexibility of *Campylobacter hyointestinalis ssp. intestinalis* in chronic diarrhea of great apes.
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 162
Title: Sunbeam: An Extensible Pipeline for Analyzing Metagenomic Sequencing Experiments
Author Block: E. L. Clarke\textsuperscript{1}, L. J. Taylor\textsuperscript{1}, C. Zhao\textsuperscript{2}, A. Connell\textsuperscript{1}, F. D. Bushman\textsuperscript{1}, K. Bittinger\textsuperscript{2};\textsuperscript{1}Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, \textsuperscript{2}Children's Hospital of Philadelphia, Philadelphia, PA.

Background: Shotgun metagenomic sequencing experiments provide functional and compositional insight into complex microbial communities. To analyze such data, a number of preprocessing and analytical steps must be performed. Many of these steps, such as quality control, adapter trimming, and phylogenetic classification, are common to many sequencing experiments. Other analyses are specific to each study. Methods: Here we introduce Sunbeam, a modular and user-extensible pipeline designed to process metagenomic sequencing data in a consistent and reproducible fashion. Sunbeam performs multiple processing steps common to many metagenomic sequencing experiments including quality control, adapter trimming, host read removal, low-complexity filtering, metagenomic classification, read assembly, and reference genome alignments. Sunbeam also includes a powerful extension framework that enables users to incorporate new analysis or processing steps easily. Results: Sunbeam installs in a single step, has no dependencies other than Linux, doesn't require administrative access, and works on most cluster computing frameworks. Sunbeam is inherently modular and will restart where it left off in case of error. To quickly and accurately filter problematic low-complexity reads in metagenomic data, we also introduce Komplexity, a rapid sequence complexity analysis tool, which identifies low complexity sequences to allow removal. The Sunbeam pipeline is well-documented, regularly updated and in routine use. We also provide a number of pre-built extensions (github.com/sunbeam-labs/). Conclusions: Sunbeam provides an easy-to-use, extensible framework for in-depth analysis of metagenomic sequencing experiments. Sunbeam ensures reproducible and consistent analyses by standardizing post-processing, analytical, and custom steps, and robust removal of problematic, low-complexity reads. Sunbeam is written in Python using the Snakemake workflow management software and is freely available at github.com/sunbeam-labs/sunbeam under the GPLv3.
Viral pathogens evolve rapidly and unpredictably, challenging the effectiveness of existing studies of viral evolution. Deep sequencing techniques detect viral mutations and diversities by sequencing a genome region multiple times. A higher coverage along a consensus sequence allows for reliable identifications of mutations among a viral population. Starting with a sufficient amount and a high purity of genomic materials is a key to obtain a high coverage consensus. One current challenge is the low virus counts in most samples, leading to sequence reads that are dominated by hosts rather than by viral pathogens. Extant enrichment methods, including virus culture and genome amplification, often introduce artificial variants or bias among sequence reads. Size-tunable-enrichment-platform (STEP), our recently developed portable technology, is constructed by aligned and functionalized carbon nanotube forests to enrich different viruses based on their sizes while removing host contaminants, e.g. host cell debris, DNA, mRNA, etc. The CNT-STEP significantly improves detection limits and virus isolation rates by at least 100 times. We integrate CNT-STEP with NGS in order to sequence unknown virus directly from field samples after enrichment. After enrichment, NGS viral reads increased from 2.9% (37,627 reads) to 90.6% (1,175,537 reads), thus corresponding to an enrichment factor of ~600, and indicating that the CNT-STEP removed most of the contamination from the host. In order to validate our new approach for real field samples, we applied a cloacal swab pool collected from five ducks during a 2012 AIV surveillance in Pennsylvania. Without any virus purification and propagation, the duck swab sample was enriched and concentrated by a CNT-STEP of 95 nm inter-tubular distance. No clogging was observed under scanning electron microscopy (SEM). NGS and de novo sequence assembly yielded 8 AIV contigs in complete lengths, but no AIV related contig was discovered in the sample without CNT-STEP enrichment. We named it “A/duck/PA/02099/2012 (H11N9)”. The H11N9 strain was further confirmed by US department of agriculture (USDA) through serological tests. This enrichment increases two orders of magnitude of sequencing coverage that dramatically enhance the sensitivity in identifying mutations. An outcome of this collaboration is the establishment of a unique method that enables close monitor of viral evolutions and a cost-effective sample preparation platform to allow for efficiency in viral deep sequencing.
Session: **Poster Session B**
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 164
Title: Whole Genome Sequencing to Track the Origin and Spread of Tuberculosis in Low Prevalence Setting of Australia

**Author Block:**

1. University of Tasmania, Hobart, AUSTRALIA
2. Trinity College, Dublin, Dublin, IRELAND
3. Royal Hobart Hospital, Hobart, AUSTRALIA
4. Launceston General Hospital, Launceston, AUSTRALIA
5. Victoria Infectious Diseases Reference Laboratory, Melbourne, AUSTRALIA.

**Abstract:**

**Background:** Tasmania is a small island state in Australia with an annual tuberculosis (TB) incidence rate of 1.7/100,000 population in 2014. A 60% drop in the current rate of TB by 2035 and a 95% drop by 2050 are required in Tasmania to meet World Health Organization’s international target of TB eradication by 2050. This study was designed to identify the source and track transmission of TB in Tasmania which is largely unknown.

**Methods:** Whole genome sequence (WGS) analyses of cultured isolates of *Mycobacterium tuberculosis* obtained from 2014 to 2016 in Tasmania was performed using Illumina Miseq at University of Tasmania. The genomic data was analyzed for single locus variation to determine phylogeny and drug resistance-conferring mutations. Genomic information was also analyzed in reference to public health surveillance records. Furthermore, *in silico* spoligotyping was performed to relate Tasmanian TB cases with publicly available isolates of International spoligotypes. Household contacts of TB cases were traced and their isolates analyzed. A cut-off of ≤5 single nucleotide polymorphism (SNP) differences between the isolates was used to define the recent transmission.

**Results:** More than 80% of TB cases in Tasmania were detected in non-Australian born individuals. Two clusters of TB were detected, one belonging to individuals originating from Nepal and other from New Zealand. Based on WGS data, isolates belonging to the largest cluster of TB in Tasmania were related to those prevalent in patient’s country of origin, Nepal. Furthermore, SNP analyses revealed Vietnam as the origin of the first case of multi-drug resistant TB in Tasmania. In addition, a human case of bovine TB reported after 40 years of its eradication from cattle in Tasmania was linked to *M. bovis* previously reported in mainland Australia.

**Conclusion:** Majority of TB cases in Tasmania have been reported in foreign-born individuals. Geographically, TB in the state had a foreign origin. Transmission of TB occurred within the members of the close community but not in a wider population.
Sequenced raw reads are available in ENA for more than 647,000 bacterial genomes. Important goals for such data may include identifying groups of genetically related bacteria in order to facilitate epidemiological tracking or in depth analyses. However, even these simple goals are difficult unless the raw data is codified.

We have developed an online tool, EnteroBase (http://enterobase.warwick.ac.uk), which provides access to genomic assemblies, genotypes and analytical tools to biologists, clinicians and epidemiologists. EnteroBase includes consistent high-resolution genotyping by core genome multi-locus sequence typing (cgMLST) schemes for Salmonella, Escherichia, Yersinia & Clostridioides and their intuitive visualization by GrapeTree (https://github.com/achtman-lab/GrapeTree) (1). Phylogenetic analyses via single nucleotide polymorphisms (SNPs) of up to 1,000 genomes are also available on-demand. An initial impression of the benefits of this approach can be found in a recent review article (2).

We are already implementing the combination of data from modern genomes with ancient DNA. EnteroBase contains more than 150,000 genomes from Salmonella and 70,000 from Escherichia. These are unprecedented troves of data on the diversity within these two genera, and the size of these databases will continue to increase dramatically over the next few years.

All read data are checked for quality, assembled and genotyped with a versioned pipeline, ensuring consistency. EnteroBase supports sharing of data within private groups of researchers as well as publishing graphical analyses and datasets for the entire global community. We are also establishing facilities to allow free download of all genomes in EnteroBase via a dedicated server. MSTree V2 and RapidNJ are implemented within GrapeTree, and can identify important clusters of related organisms among 100,000 genomes based on cgMLST. However, we are already preparing for the future that will encompass orders of magnitude more genomes, by developing hierarchical clustering, which will provide persistent and scalable designations, as a general tool for microbial genomics.

Reference List
Session: Poster Session B
Time: Tuesday, September 25, 2018, 2:00 pm - 3:30 pm
Poster Board #: 166
Title: Comparison of Genomic Analysis Methods for Investigation of a Legionellosis Cluster in New York City, October 2017
1CDC, Atlanta, GA, 2NYS Wadsworth Center, Albany, NY, 3NYC PHL, NYC, NY, 4NYS Wadsworth, Albany, NY.

Abstract:
Legionellosis is caused by exposure to Legionella species found in water; symptoms range from a mild influenza-like illness to a serious and sometimes fatal form of pneumonia. In the United States, ~79% of cases are associated with Legionella pneumophila serogroup 1 (Lp1). Legionella is a growing public health concern in the country; disease incidence has nearly quadrupled since 2000 with several large high-profile outbreaks in the recent years, including in New York City. During October 1-14, 2017, 15 cases of legionellosis were confirmed in a <0.75 km radius in Flushing, Queens, NY. We used 2 comparative genomic methods to characterize environmental isolates collected during the investigation and compare them with circulating strains to assess diversity of Lp1. During the environmental investigation, 55 epidemiologically linked cooling towers and water fountains were sampled and tested by culture-based methods and by real-time PCR for presence of Legionella species, Legionella pneumophilia, and Legionella pneumophila serogroup 1 (Lp1). No clinical isolates were recovered from 6 sputum specimens obtained from patients, but 13 Legionella species isolates and 5 Lp1 isolates were recovered from environmental sources. Whole-genome sequencing (WGS) was performed on all 5 Lp1 isolates and single nucleotide polymorphism (SNP) and multilocus sequence typing (wgMLST) were carried out for in-depth molecular characterization of circulating strains. SNP analysis was used to compare isolates recovered during the investigation with historical Lp isolates from New York State (NYS). One isolate matched two unrelated clinical isolates from NYS with 24-25 SNPs differences, whereas, remaining isolates were closely related to each other and environmental isolates previously recovered during the 2015 South Bronx outbreak, which indicates persistence of this strain in NYC. In silico sequence-based typing revealed that 4 isolates were sequence type (ST) 1400 that has been found only in NY and 2 of these isolates shared a high degree of similarity (>99% allele identity) with a clinical isolate from 2009 recovered in NYC. WGS has provided additional resolution to outbreak investigations. In this investigation, two different genomic sequence analysis methods were used with comparable results. Both methods were able to distinguish and separate isolates based on their relatedness. We conclude that particular Legionella strains could be endemic and persistent in NYC based on similarity of strains and ST unique to NY. Additionally, we detected diversity of potential disease-causing strains in cooling towers when compared with strains commonly found in the United States. Our investigation showcases how WGS is crucial in outbreak investigations and highlights need for obtaining clinical isolates from patients with legionellosis to identify disease sources to prevent additional exposures.
Prevalence and Serovar Diversity of *Salmonella* spp. in Primary Agricultural Horticultural Fruit Production Environments

Author: L. Chidamba, L. L. Korsten, A. Gomba; University of Pretoria, Pretoria, SOUTH AFRICA.

Abstract: Increases in foodborne disease outbreaks associated with fresh produce have necessitated the need to identify potential sources of microbial contamination in produce and agricultural environments. The present study evaluated *Salmonella* prevalence and serovar diversity in fruit (225), water (140) and surface (126) samples, from three commercial farms and associated packhouses, located in different farming regions in South Africa. Fruit and water samples were collected from both orchards and packhouses, while surface samples were collected from conveyer belts and hands of packhouse employees. *Salmonella* was detected in 26 of the 491 (5.3%) samples. Environmental samples (water and surfaces) recorded a slightly higher proportion (3.1%; 15/491) of positive samples compared to fruit samples (2.2%; 11/491). *Salmonella* was not detected on employee hands and river water samples. A total of 263 *Salmonella* cultures were isolated from the 26 from positive samples by standard culture methods, preliminarily identified through matrix-assisted laser desorption ionisation-time of flight mass spectroscopy (MALDI-TOF MS) and API 20E, and confirmed by invA gene. Of the 39 representative isolates serotyped the serovars Muenchen (33.3%), Typhimurium (30.8%), Heidelberg (20.5%), Bsilla (7.7%), *Salmonella* subspecies IIb: 17: r: z (5.1%) and one untypable strain were identified. Most samples had multiple serovars with orchard water form one site recording the highest serovar diversity (4 serovars). Our findings show the potential of agricultural fruit production environments to act as reservoirs of clinically important *Salmonella* serovars.
Session: **Poster Session B**  
**Time:** Tuesday, September 25, 2018, 2:00 pm - 3:30 pm  
**Poster Board #:** 168  
**Title:** Applying Bioinformatics Pipelines to Reconstruct Bacterial Genomes from Human Faecal Metagenomes  
**Author Block:** F. M. Mobegi¹, L. E. Leong¹, B. Ramadass², E. Mortimer³, M. J. Manary⁴, D. H. Alpers⁵, G. P. Young⁶, B. S. Ramakrishna⁶, G. B. Rogers¹;  
¹SAHMRI, Adelaide, AUSTRALIA, ²All India Institute of Medical Sciences, Odisha, ³Flinders University, Bedford Park, AUSTRALIA, ⁴Washington University in St. Louis, St Louis, MO, ⁵Washington University in St. Louis, St. Louis, MO, ⁶SRM Institutes for Medical Science, Chennai, INDIA.  
**Introduction:** Advances in metagenomics and computational methods, together with reductions in sequencing costs, have aided culture-independent studies of complex microbial systems. Metagenomics-based analysis has primarily focused on assessments of microbiome diversity and functional capacities. However, deep metagenomics sequencing also allows the reconstruction and exploration of draft microbial genomes. This approach is particularly powerful in relation to culture-refractory species. We aimed to retrieve high-quality draft bacterial genomes from faecal metagenomes, generated from pre-school children in Tamil Nadu, India.  
**Methods:** Shotgun metagenomic sequencing was performed on longitudinal stool sample collections from three stunted and three non-stunted adolescents who were enrolled in a starch supplementation study. On average, 261 million high-quality reads were obtained from each sample. Using IDBA and CD-HIT, the reads were *de novo* assembled into contigs and dereplicated to remove redundant sequences respectively. Individual sample reads were then mapped to the non-redundant contigs using Burrow-Wheeler Aligner (BWA), and the resulting BAM files sorted and indexed using Samtools. MetaBAT, with all the five preset parameters, and a depth file of each BAM file, was used to bin contigs, as previously described¹. Resulting bins, which represent draft genomes, were assessed for completeness and purity and refined using checkM and RefineM. Taxonomic assignments for the drafts were confirmed using Kraken.  
**Results:** Based on deep metagenomic sequencing we reconstructed near-complete genomes of 114 bacterial taxa with high genome quality (completeness ≥70%, contamination ≤10%). Approximately 93% of all recovered genomes represented fermentative commensal species. Although the reconstructed genomes displayed notable consistency with their type-strains in the core genome composition, some selected culture-refractory anaerobic bacteria revealed significant differences with their type-strain counterparts in the accessory genome. In monosaccharides metabolism, for example, reconstructed *A. muciniphila* has genes needed to utilise d-galacturonate and d-glucuronate, which are absent in the type-strain *A. muciniphila* (ATCCBAA-835). In contrast, the ATCC strain has genes for fructose utilisation that are absent in our genome. These differences might reflect characteristics of local diet.  
**Conclusion:** We demonstrated the successful recovery of draft bacterial genomes from faecal metagenomes and their comparison to type-strains. This ability to construct bacterial genomes directly from metagenomes is valuable in allowing the analysis of culture-refractory taxa, and is likely to be particularly important in contexts where advanced culture techniques are unavailable. It also provides a means to mine existing published datasets.  
**Literature**  
**Background:** Public health laboratory surveillance systems have historically relied on two types of assays to detect pathogens, which include i) profiling the organism-of-interest via morphological traits (microscopy), metabolic capability (culture), and/or molecular subtyping, or ii) surveying the environment of the organism-of-interest through serology. With decreasing costs, next generation sequencing (NGS) has emerged as a surveillance tool that potentially provides a standardized protocol across groups of microorganisms and finer resolution than subtyping. Unfortunately, the implementation of NGS and bioinformatics analyses in state laboratories has remained challenging.

**Methods:** Massachusetts State Public Health Laboratory (MA SPHL) was funded as the bioinformatics leader laboratory in 2018 for the New England area, which includes CT, MA, ME, NH, NY, RI, VT, NYC, and NJ. In order to determine the status of bioinformatics for these laboratories, MA SPHL hosted a series of 6 calls in partnership with the Broad Institute of MIT and Harvard. The first 3 calls reviewed each state’s sequencing and bioinformatics infrastructure, while the latter 3 calls consisted of demonstrations of cloud computing through Amazon Web Services (AWS) and Google Compute Engine.

**Results:** Surveys and calls revealed that most states had MiSeq sequencing capability, and participated in CDC programs such as PulseNet, National Antimicrobial Resistance Monitoring System (NARMS), Global Health Outbreak and Surveillance Technology (GHOST), and CaliciNet. Wadsworth NY was an outlier with both sequencing and bioinformatics cores, along with active assay and pipeline development for organisms outside those projects (for e.g, adenovirus, mumps, and zika). Difficulties in setting up bioinformatics infrastructure stemmed from information technology (IT) resistance for Linux or Cloud support, little funding for bioinformatics staff, and lack of data policies for sequencing data.

**Conclusions:** Most New England states have obtained the ability to sequence by participating in CDC programs, but rely on the CDC or outside collaborators for bioinformatics analyses. Each state will likely require its own unique solution due to differing state laws and governance, which in turn shapes state IT departments and prevents them from emulating CDC’s compute model. To provide future guidance, we are currently drafting a “Bioinformatics Implementation Guide”, which will include considerations for hiring bioinformaticians, finding compute hardware, and working with IT.