ontrol Number: 2022-A-32-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: A Universal "Day Zero" infectious disease test leveraging CRISPR based host depletion and meta genomic shotgun sequencing

Author Block: **K. Brown**;
Jumpcode Genomics, San Diego, CA.

Abstract Body: The lack of preparedness for detecting the highly infectious SARS-CoV-2 pathogen — the pathogen responsible for the COVID-19 disease — caused enormous harm to the public health, the economy and society as a whole. It took ~60 days for the first RT-PCR tests for SARS-CoV-2 infection developed by the United States (US) Centers for Disease Control (CDC) to be made available. It then took >270 days to deploy 800,000 of these tests at a time when the estimated actual testing needs required over 6 million tests per day. Testing was therefore limited to only individuals with symptoms or individuals in close contact with confirmed positive cases. Testing strategies that can be deployed on a population scale at ‘day zero’ - i.e., at the time of the first reported case - are needed. Next Generation Sequencing (NGS) has day zero capabilities with the potential to enable feasible and very broad large-scale testing strategies. We show that the detection sensitivity of SARS-CoV-2 via NGS is equivalent to RT-PCR detection if relevant samples are depleted of abundant sequences that don’t contribute to pathogen detection or host response. In addition, we show that the proposed strategy can also be used for variant strain typing, co-infection detection, and individual human host response assessment - all in a single workflow using existing open-source analysis pipelines. The NGS framework we describe is pathogen agnostic, and therefore has the potential to radically transform how both very large-scale pandemic response and focused clinical infectious disease testing are pursued in the future.

---PAGEBREAK---

Control Number: 2022-A-69-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: EDGE COVID-19: A Web Platform to generate submission-ready genomes, wastewater data, and genome recombinant analysis from SARS-CoV-2sequencing efforts.

Author Block: **C. Lo**, P. Li, M. Shakya, B. Hu, P. Chain;
Los Alamos National Laboratory, Los Alamos, NM.

Abstract Body: During the current outbreak, genomics has become an essential toolfor surveilling infectious disease outbreaks. A wide range of strategies andtechniques for sequencing and processing SARS-CoV-2 genomes are being used bylaboratories worldwide. These methods are quite different, and especially forcomputational processing, sometimes ad hoc. A standardized, well tested, andaccessible tool for consensus genome sequence determination, particularly foroutbreaks such as the ongoing COVID-19 pandemic, is critical to provide a solidgenomic basis for epidemiological analyses and well-informed decision making.Additionally, a tool that goes beyond consensus genome generation and canidentify recombinants and mixed infection is warranted based on currentevolutionary track of the virus. Here, we have developed a bioinformaticworkflow called EDGE COVID-19 (EC-19) that is capable of generating consensusgenomes of SARS-CoV-2 and further perform pertinent downstream analyses. EC-19 accommodates sequencing data generated with either the Illumina or OxfordNanopore or Pacbio platforms. Furthermore, using an intuitive web-basedinterface, this workflow automates SARS-CoV-2 reference-based genomeassembly, variant calling, lineage determination, wastewater sample lineageabundance estimation, genome recombinant analysis and provides the ability tosubmit the consensus sequence and necessary metadata to GenBank, GISAID, andINSDC raw data repositories.
Availability: https://edge-covid19.edgebioinformatics.org

---PAGEBREAK---

Control Number: 2022-A-73-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: refMLST: Reference-based Multilocus Sequence Typing Enables Universal, Decentralized Bacterial Typing

Author Block: I. Chandrakumar, **S. D. Chorlton**;
BugSeq Bioinformatics Inc., Vancouver, BC, CANADA.

Abstract Body:

Background

Despite efforts to standardize genomic outbreak analyses, both single nucleotide variant (SNV) and gene-by-gene (eg. cgMLST/wgMLST) approaches continue to be used. Multiple comparisons have demonstrated these approaches perform largely equivalently; however, SNV approaches are limited by challenges of data sharing and genetic recombination adjustment, while gene-by-gene approaches are limited by the need for a stable, curated, centrally-hosted scheme, limiting their utility across all bacteria.

Methods

We present refMLST, a tool for reference-based multilocus sequence typing of bacteria. refMLST functions by building a scheme from a single, “reference” genome annotation, and therefore functions across all bacterial species with a genome sequence available. We use NCBI’s reference genome annotation for each species, but any GenBank formatted file may be used. refMLST excludes mobile elements such as plasmids from the scheme. Similar to gene-by-gene analysis, scheme genes are then located in the query genome. Instead of a comparison to a pre-computed scheme to identify allele names, allele sequences are hashed to yield an integer identifier for the allele of each gene. Bacterial genomes are compared against each other by comparing the hash at each gene in the scheme, and a distance matrix is computed based on allele distances.

Results

We applied refMLST to 1263 Salmonella enterica isolates with publicly available sequencing data, and compared results to chewieSnake, a recently published, decentralized, hash-based cgMLST tool. We find the overall correlation coefficient between refMLST and chewieSnake cgMLST allele distances to be 0.95. At distances less than 100 refMLST alleles, refMLST distances are on average 20 (SD: 10) alleles greater than chewieSnake, indicating an enhanced ability to resolve closely related isolates. Increased allele distances are explained by examination of a greater number of loci with refMLST. Accounting for differences in allele distances between tools with linear regression, we find outbreak clusters at comparable allele distances (refMLST: 20 alleles, chewieSnake: 10 alleles) to be highly correlated with an Adjusted Rand Index of 0.92.

Conclusion

refMLST combines the advantages of SNV and gene-by-gene approaches to enable decentralized, reproducible bacterial typing without suffering from each approach’s limitations. refMLST has already been applied to hundreds of genomes across public health laboratories and is freely available for academic use at https://bugseq.com/academic.

---PAGEBREAK---

Control Number: 2022-A-83-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Prediction of Protein-encoding Gene Content in *Escherichia coli* Genomes

Author Block: **J. J. Davis**;
Argonne National Laboratory, Lemont, IL.

Abstract Body: Having the ability to predict the protein-encoding gene content of a genome is of vital importance for a variety of bioinformatic tasks including binning genomes from metagenomic sequences, estimating genome completeness, and assessing risk due to the presence of antimicrobial resistance and other virulence genes. In this study, as a proof of concept, we built machine learning classifiers for predicting the presence or absence of the variable genes occurring in 10-90% of all publicly available high-quality *E. coli* genomes. The PATRIC genus-specific protein families were used to define orthologs across the set of genomes, and a single binary classifier was built for predicting the presence or absence of each family in each genome. Each model was built using the nucleotide k-mers from a set of 100 conserved genes as features. The resulting set of 3,259 XGBoost classifiers had an average macro F1 score of 0.907 ± 0.905-0.910 (± 95% confidence interval over 5 folds). Models predicting the presence or absence of genes that were well annotated, either occurring in subsystems, or having full EC numbers, had significantly higher F1 scores than those that did not, and models for genes with annotations involved with horizontal gene transfer, including those with the terms “plasmid” and “conjugation” had significantly lower F1 scores. We show that the F1 scores are stable across MLSTs, and that the trend can be recapitulated through sampling with a smaller numbers of core genes. Furthermore, we evaluate the extensibility of the models by classifying a holdout set of 422 diverse *E. coli* genomes that were isolated from freshwater environmental sources. Overall, this study provides a framework for predicting gene content from a limited amount input sequence data.

---PAGEBREAK---

Control Number: 2022-A-110-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Distributed Gene Embeddings for the Classification of Candidate Divergent Antimicrobial Resistance Genes

Author Block: **J. T. Lewandowski**, R. G. Beiko;
Dalhousie University, Halifax, NS, CANADA.

Abstract Body: **Background:** Antimicrobial resistance (AMR) is a critical global health crisis, and methods to identify emerging AMR genes are essential to monitor and clinically manage its propagation. While homology is the primary line of evidence for predicting gene function, conserved gene order can also provide valuable information in instances where gene sequences diverge from known AMR conferring genes. Previous methods used for the classification of gene function have included gene order analysis, gene embeddings, and network representations of genomic data. However, these approaches in isolation may not integrate enough genomic context information to rule out false positives. **Objective:** In this work, we aim to train classifiers on gene embeddings to predict candidate divergent AMR genes. Here, an embedding is a vector representation of a given gene that encodes the occurrence of surrounding genes in its neighborhood across many genomes, allowing us to maximize the gene order context when making functional predictions. To generate such representations, we use gene order data in combination with G2Vec, a deep-learning method that uses network-based learning to construct gene embeddings and which has been successfully used to identify cancer prognostic genes. We posit that divergent AMR genes can be identified if they receive a high probability score from classifiers trained on the top gene embeddings of known AMR-conferring genes and are supported by the presence of other AMR genes in close proximity. **Methods:** We used G2Vec to obtain distributed gene representations for every gene present in the pangenomes of Enterococcus faecium, Escherichia coli, Salmonella enterica, Mycobacterium tuberculosis, and Klebsiella pneumoniae and identified the most representative embeddings for both the AMR-conferring and non-AMR-conferring sets. We then trained random forest, support vector machine, and XGBoost classifiers on these embeddings to predict probabilities for candidate AMR-conferring genes. **Results:** t-SNE visualizations of the top AMR-conferring and non-AMR-conferring representations for all species showed separable and tightly packed clusters, suggesting that they shared similar conservation characteristics. Our experimental results indicate that the optimized Random Forest classifier performed best, having achieved a 90.62% ROC-AUC score which signifies a strong ability to differentiate between the two sets. The trained model was used to predict the class probabilities of possible divergent AMR genes from all species and identify the top 10 highest supported candidates which were verified in terms of their functional annotations and top homologous protein hits against the Comprehensive Antibiotic Resistance Database (CARD). **Conclusions:** By incorporating existing knowledge about gene neighborhoods, the embedding-based approach of G2Vec was able to make accurate predictions in cases where the evidence from homology was less clear.

---PAGEBREAK---

Control Number: 2022-A-115-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Alignment-free Recombination Detection Using Genomic Database Distributions of Exact Protein Matches

Author Block: **A. M. Moustafa**, A. Lal, P. J. Planet;
University of Pennsylvania, Philadelphia, PA.

Abstract Body: Genomic recombination in microbial populations is a critical generator of biological diversity, and it plays an important role in adaptation to new environments, hosts, and niches. Recombination detection has previously relied on alignment of genomic sequences and phylogenetic or comparative techniques that are extremely computationally expensive, especially with vast increases in available whole genome sequences. Here we present a database-driven technique that is alignment-free, does not rely on phylogeny or sequence similarity, and can be used on single genomes. The tool, Redcarpet (**Re**combination **D**etection using **C**omparative **A**nalysis of **R**egional **P**atterns of **E**xact Match **T**argets) combines the output from our recently developed WhatsGNU algorithm with a MinHash technique. Redcarpet takes in a single query genome, and for each protein determines the set of genomes in a database that contain an exact protein sequence match. It then computes the similarity between genome sets for all pairwise gene comparisons in the genome. This operation is predicated on the idea that identical genes are more likely to appear in the same set of genomes if they share a common evolutionary history. Redcarpet outputs a pairwise, genome-set similarity matrix that can be visualized as a 2-D heatmap ordered by the gene location on the chromosome. The heatmap provides a visual tool for identifying recombination tracts. We use probabilistic changepoint analysis on this matrix to identify likely recombination break points. The genome sets determined by Redcarpet can also be used to identify the likely origin of genomic segments in known genome subgroupings (e.g., clonal complexes), and to identify a genomic "core" for subsequent phylogenetic analysis. We tested Redcarpet on simulated data and also on known examples of large-scale recombination events in *Staphylococcus aureus* and *Klebsiella pneumoniae*. Redcarpet can be used to rapidly identify recombination tracts in any species that has a large database of genomic sequences.

---PAGEBREAK---

Control Number: 2022-A-119-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Pandemic-scale phylogenetic analysis and phylogenetic dashboard in the European Bioinformatics Institute's COVID-19 Data Portal

Author Block: **J. Szarvas**, M. Koliba, F. M. Aarestrup;
Technical University of Denmark, Kgs. Lyngby, DENMARK.

Abstract Body: Sample sequencing efforts for the SARS-CoV-2 pandemic generated large quantity of sequences deposited in the public nucleotide databases. Utilizing this genomic data, together with the minimum metadata available, in phylogenetic analysis could elucidate the evolution of the virus close to real time.

We set up and maintain an instance of a genetic distance-based, large-scale phylogenetics analysis pipeline for processing public SARS-CoV-2 consensus sequences in the European Nucleotide Archive. A phylogenetic visualization dashboard web-component (PhyloDash) was also developed. A web-app interactively displaying the results is available through the COVID-19 Data Portal maintained by the European Bioinformatics Institute. The dashboard contains the most recent WHO region specific phylogenetic trees, inferred on the hundreds of thousands of publicly available SARS-CoV-2 consensus sequences. In addition, there is metadata, pango-lineage information, and S and N gene mutations listed for each sample, that can be used to search and filter the trees.

---PAGEBREAK---

Control Number: 2022-A-122-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Dynamic Neighbor-Joining: Scaling Neighbor-Joining to Millions of Taxa with Dynamic Programming

Author Block:
P. Clausen;
DTU, Kgs. Lyngby, DENMARK.

Abstract Body:
Dynamic Neighbor-Joining: Scaling Neighbor-Joining to Millions of Taxa with Dynamic Programming*Philip T.L.C. Clausen*[11]*Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, 2800 Kgs Lyngby, Denmark*
AbstractThe Neighbor-Joining algorithm is a widely used method to perform hierarchical clustering, and forms the basis for phylogenetic reconstruction in several bioinformatic pipelines. With a runtime complexity of O(n^3) the Neighbor-Joining algorithm is considered to be a computational efficient algorithm, however, it does not scale well for datasets exceeding several thousand taxa (> 100 000), as is more and more commonly seen today. Although optimizations to the canonical Neighbor-Joining algorithm have been proposed, these optimizations are achieved through approximations or extensive memory usage, which is infeasible for large datasets.Here I present Dynamic Neighbor-Joining, which optimize the canonical Neighbor-Joining method to scale to millions of taxa, without increasing the memory requirements as only one lower triangular distance matrix is stored. Dynamic Neighbor-Joining outperform the current gold standard methods to construct exact and approximate Neighbor-Joining trees, while Dynamic Neighbor-Joining is guaranteed to produce exact Neighbor-Joining trees.As the memory requirements remains at O(n^2), the implementation has been performed such that the distance precision can be set from eight to one byte(s) using runtime options. This allows users to construct massive trees of rapidly evolving taxa, such as seen for SARS-CoV-2, where only short distances are observed due to the massive surveillance performed worldwide.**Availability and requirements**Repository: https://bitbucket.org/genomicepidemiology/ccphylo.git License: Apache-2.0Operating system(s): Unix based systems.Programming language: C.Other requirements: zlib development files.

---PAGEBREAK---

Control Number: 2022-A-142-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: NCBI Pathogen Detection surveillance enables discovery of novel class A and class D beta-lactamase genes in Enterobacterales

Author Block:
M. Feldgarden, V. Brover, B. Fedorov, D. H. Haft, A. B. Prasad, W. Klimke;
NCBI/NLM/NIH, Bethesda, MD.

Abstract Body:
The spread of antimicrobial resistance (AMR) genes is a major global public health problem, and open and freely accessible data is critical to meeting this threat. The National Center for Biotechnology Information (NCBI) Pathogen Detection system develops and maintains the AMRFinderPlus software and databases to identify anti-microbial resistance (AMR) and other genes important for surveillance. As part of the Pathogen Detection system, NCBI runs AMRFinderPlus on the over 1,000,000 publicly available bacterial genomic sequences submitted by public health agencies, researchers, hospitals, and others, and provides then public reports for outbreak investigation and AMR gene surveillance.

Here we use the NCBI Pathogen Detection web interfaces that report AMRFinderPlus results to identify novel putative class A and class D beta-lactamase genes in Enterobacterales isolates. AMRFinderPlus utilizes a combination of protein BLAST-based similarity searches, hidden Markov models (HMMs), and a hierarchy of genes and families to identify and infer function for both known and undescribed AMR genes. NCBI curators have organized many gene families into higher-level nodes containing multiple homologous genes: putative AMR genes belonging these nodes are identified using highly curated HMMs. The combination of HMMs and a hierarchy built on existing beta-lactamases can be used to identify novel potential beta-lactamases and carbapenemases. Specifically, we determined which proteins were identified solely through HMM hits to a class A carbapenemase HMM or to a class D beta-lactamase HMM.

We identified and describe two putative novel class A carbapenemases. One putative carbapenemase is a KPC family variant with a ten amino-acid deletion of unknown function, and which has been found independently on a plasmid in a second bacterial species. The other putative carbapenemase was found in two separate *Enterobacter cloacae* isolates. We also found four novel class D beta-lactamases of unknown function, one of which could be localized to a plasmid occurring in multiple species, and two of which appear to be associated with integron-related genes.

We align and place these genes in phylogenetic context and suggest possible function based on their relationship to known members of these closely related gene families. These data demonstrate how NCBI’s Pathogen Detection website can be used to identify high confidence novel candidate AMR genes of possible public health importance and suggest possible resistance spectra for further experimental characterization.

---PAGEBREAK---

Control Number: 2022-A-143-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: Democratizing access to microbial bioinformatics tools

Author Block:
F. Thibaud-Nissen, F. the NCBI prokaryotic genome annotation team;
NCBI/NLM/NIH/DHHS, Bethesda, MD.

Abstract Body:
Gene annotation is a prerequisite to many genomics analyses and is essential for unlocking the potential of microbial sequencing. Our goal is to offer bioinformatics tools that are easy to use and produce a quality product on bacterial and archaeal genomic reads or genomes. RAPT, the Read assembly and Annotation Pipeline Tool (https://www.ncbi.nlm.nih.gov/rapt) is an easy-to-use web service for the assembly and annotation of prokaryotic genomic reads. It provides an all-in-one solution for users with no bioinformatics expertise or no access to computing resources. The user provides an SRA accession or short read fastq files for a sequenced genome and receives the corresponding annotated assembly a few hours later. RAPT is the latest in a suite of tools developed by NCBI to analyze prokaryotic genomes. It utilizes SKESA for assembling short reads the user provides on input into an assembly, and PGAP, the annotation pipeline used for producing annotation on RefSeq genomes, to annotate the assembly. PGAP is a mature pipeline that incorporates expert-curated evidence and is used for maintaining the annotation of over 230,000 RefSeq genomes. It is also available as a Docker container that can be executed outside of NCBI on an individual computer, a compute farm, or in a cloud environment from an intuitive command-line interface (see https://github.com/ncbi/pgap). We will present how recent additions to the PGAP and RAPT workflows provide a richer gene annotation that includes Gene Ontology terms, and an assessment of genome quality and completeness.

---PAGEBREAK---

Control Number: 2022-A-152-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: A Method for Automated Designation of Viral Lineages

Author Block:
J. D. McBroome, J. Lyda, R. Corbett-Detig;
University of California Santa Cruz, Santa Cruz, CA.

Abstract Body:
The public health response to the COVID-19 pandemic has demanded the effective integration of phylogenetic and epidemiological analyses on vast quantities of sequencing data. One critical element of this integration is effective viral lineage identification and tracking. The Pangolin lineage system is the gold standard for SARS-CoV-2; however, it is based on crowd-sourced proposals stemming from manual scrutiny of the global phylogenetic tree. This approach is dependent on a large and active epidemiological community and is vulnerable to regional and personal bias. In this work we present an automated alternative approach that divides a phylogenetic tree into lineages based on relative genotype representation. In brief, we identify nodes within the phylogenetic tree that represent a substantial portion of the total genotype of descendents, compared to their ancestral lineage or the root. We designate these nodes as lineage roots and their descendents as members of that lineage; this is repeated until no nodes confer a relative increase of genotype representation under the user’s parameters. Our method is efficient on extremely large phylogenetic trees and produces similar results to existing pangolin lineage designations. Consistent and timely lineage designation is a key component of pandemic preparedness, and this approach can serve to assist researchers in the effective and rapid identification of epidemiologically relevant lineages going forward.

---PAGEBREAK---

Control Number: 2022-A-196-NGS

Session Title: **Pipe Dreams II (Lightning Talks)**

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

Publishing Title: A Scalable Suite of PCR design and evaluation tools

Author Block:
P. Davis;
MRIGlobal, Gaithersburg, MD.

Abstract Body:
Sequence data is facilitating a rapid shift in the surveillance of emerging pathogens of both agricultural and human health significance. The utilization of molecular techniques reliant on PCR has grown to include both quantitative PCR assays for diagnostics and amplicon sequencing for molecular epidemiology. Despite advances in data availability and demand for these molecular solutions, primer design is labor intensive and often requires specific knowledge developed from narrow, or data starved experiments. Also, as demonstrated through the Spike Gene Target Fault (SGFT) signal in the current SARS-CoV-2 pandemic, consistent PCR assay evaluation is critical to assessing performance of diagnositics. We have developed an automated PCR primer design pipeline for the generation of candidate primer sets constrained by design specifications on

either curated user sequence data, or taxonomy. To select candidate primer sets, or to evaluate the performance of existing primer sets, we have developed a suite of tools for analyzing different characteristics of PCR based assays, either probe or non-probe-based, for inclusivity, exclusivity and primer interactions. As a demonstration of the need for more rapid tools for development and evaluation we examine to emerging pathogen examples.

The rapid spread of African Swine Fever Virus (ASFV) in 2019 posed a serious threat to global domesticated pig farming. The World Organisation for Animal Health, formerly the Office International des Epizooties (OIE), published in their Terrestrial Manual a chapter on ASFV, including previous published primer sets for the detection of ASFV via real-time PCR. This assay was subsequently discovered to inefficiently detect ASFV, and in silico PCR analysis revealed none of the assays resulted in perfect primer matches for the available sequences in 2019. We then evaluated primers from three other publications leading up to 2019 and discovered similar issues with a single primer set from Luo et al. 2016, maintaining perfect matching. Mismatches in primer sets have been shown to diminish assay sensitivity, in a position and assay parameter dependent manner. To demonstrate how readily better solutions are potentially available, we used our pipeline to generate 96 candidate primer pairs with similar thermodynamic and product size characteristics as the OIE assays that perform better on available sequence data in 2019.

At this moment, Monkeypox virus is spreading rapidly across mulptiple continents. Similar to the ASFV situation in 2019, we find issues with the current CDC assay test procedures. In fact, of the 402 available complete Monkeypox genomes available on NCBI Virus as of July 27th 2022, only 38 (9%) are expected to have exact primer matches in both the forward and reverse primer alignments. Effects on potential assay sensitivity are not evaluated as far as we are aware.

---PAGEBREAK---