– Poster Presentations –

#1: Annotation of Long Metagenomic Sequences, Prokaryotic and Eukaryotic

Presenting Author: **Alexandre Lomsadze**
Email: alexl@gatech.edu

Complete Author List: Alexandre Lomsadze (1,2); Liexiao Ding (5); Wenhan Zhu (2); Mark Borodovsky (1,2,3,4)

(1) Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA
(2) Center for Bioinformatics and Computational Genomics, Georgia Tech, Atlanta, GA
(3) School of Computational Science and Engineering, Georgia Tech, Atlanta, GA
(4) Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia
(5) H. Milton Stewart School of Industrial & Systems Engineering, Georgia Tech, Atlanta, GA

Abstract:

Assembly of metagenomic reads into longer contigs becomes a standard step in metagenomics studies. The increase in the contig length presents an opportunity to improve metagenome annotation. We describe a new version of the gene prediction algorithm, MetaGeneMark-2, that assigns locally optimized parameters to each ORF in a metagenomic contig and improves prediction accuracy in contigs with heterogeneous GC composition. Also, the longer contigs of eukaryotic metagenomics sequences (fungi and protists) provide more data for learning the parameters of a eukaryotic gene finder. The new algorithm selects a model from a set of kingdom specific heuristic models pre-built for a wide range of GC contents.

#2: Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale

Presenting Author: **Karl Gemayel**
Email: karl@gatech.edu

Complete Author List: Alexandre Lomsadze* (1,2); Karl Gemayel* (3); Shiyuyun Tang* (5); Mark Borodovsky (1,2,3,4); *joint first authors

(1) Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA
(2) Center for Bioinformatics and Computational Genomics, Georgia Tech, Atlanta, GA
(3) School of Computational Science and Engineering Georgia Tech, Atlanta, GA
(4) Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia
(5) School of Biological Sciences Georgia Tech, Atlanta, GA

Abstract:

GeneMarkS-2, a new ab initio gene finder, aims to improve prediction of species-specific (native) genes, as well as difficult-to-detect genes that differ in composition from the native genes. We introduce an array of pre-computed heuristic models that compete with the iteratively learned native model for the best fit within genomic neighborhoods that deviate in nucleotide composition from the genomic mainstream. Also, in the process of self-training, GeneMarkS-2 identifies distinct sequence patterns controlling transcription and translation. We assessed the accuracy of current state-of-the-art gene prediction tools on test sets of genes validated by proteomics experiments, by COG annotation, as well as by protein N-terminal sequencing. We observed that, on average, GeneMarkS-2 shows a higher precision in all accuracy measures. Screening of ~5,000 representative prokaryotic genomes revealed widespread leaderless transcription, not only in archaeal domain where it was originally discovered, but in bacterial domain as well. Furthermore, the RBS sites of the species with prevalent ‘leadered’ transcription may frequently exhibit consensus patterns different from the conventional ones of Shine-Dalgarno. GeneMarkS-2 distinguishes leaderless and leadered transcription and reveals the prevalence of one or the other, thus, making classification of prokaryotic genomes into five groups with distinct sequence patterns around gene starts. Some of the observed patterns are apparently related to yet poorly characterized mechanisms of translation initiation.

#3: Predicting Health States from Cystic Fibrosis Lung Microbiome Composition

Presenting Author: **Conan Y. Zhao**
Email: czhao98@gatech.edu

Complete Author List: Conan Y. Zhao (1); Karan A. Kapuria (1); Yiqi Hao (1); John Varga (2); Sam P. Brown (1); Joanna B. Goldberg (2)

(1) Georgia Institute of Technology, School of Biological Science
(2) Emory University, Department of Pediatrics

Abstract:

Many studies show that patient health status is directly related to their microbiome composition. However, inferring these relations and developing a predictive model for health status is difficult due to the complex dynamics of microbial communities. Many algorithms infer mechanistic models by fitting high-resolution longitudinal data to generalized Lotka-Volterra equations. However, in a patient health context such data can be difficult to obtain. Other methods avoid the need for temporal resolution by calculating correlational information from 16S pyrosequencing snapshot data. Predictions based on microbe correlations, however, can differ greatly from those based on interaction data. We present a novel analysis pipeline for predicting patient health metrics using 16S compositional data. We apply machine learning techniques to fit general Lotka-Volterra models using simulated microbiome data as well as cystic fibrosis (CF) lung microbiome data and metadata from a cohort of 77 patients. We predict community metrics such as stability and antibiotic susceptibility and compare our results against other commonly used correlational analysis techniques.

#4: Staphopia: an analysis pipeline and Application Programming Interface focused on *Staphylococcus aureus*.

Presenting Author: **Robert A. Petit III**
Email: robert.petit@emory.edu

Complete Author List: Robert A. Petit III (1); Timothy D. Read (1,2)

(1) Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA, USA
(2) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

Abstract:

Rapid low-cost sequencing of clinically-important bacterial pathogens has generated thousands of publicly available datasets and many hundreds of thousands more will undoubtedly soon be generated. Analyzing these genomes and extracting relevant information for each pathogen and the associated clinical phenotypes requires not only resources and bioinformatic skills but domain knowledge on the nuances of the organism. We have created an analysis pipeline and API focused on Staphylococcus aureus, which is not only a common human commensal but is also of important public health interest, with MRSA (methicillin-resistant S. aureus) a major antibiotic-resistant hospital pathogen. Staphopia can be used both for basic science studies (e.g. patterns of evolution) but also potentially as a platform for rapid clinical diagnostics. Written in Python, Staphopia's analysis pipeline consists of submodules running open-source tools managed by a pipeline manager. It accepts raw FASTQ reads as an input, which undergo quality control filtration, error correction and reduction to a maximum of 100x coverage. This data reduction is advantageous for load management when processing thousands of genomes. Using preprocessed reads the pipeline branches off into de novo assembly-based analysis and mapping-based analysis. Modules running species-specific analyses such as antibiotic resistance profiling and multi-locus sequence type (MLST), use the contigs. Genes are annotated from the contigs using PROKKA and the UniProt database. Mapping is used to to call all variants (SNPs and InDels) against a single reference chromosome (S. aureus N315). From the processed reads, 31-mers are counted for each input sample. Depending on the size of the input file, analysis is completed between 20-60 minutes. With Staphopia's web application, built using the Django web-framework, analysis results from each genome are stored within a PostgreSQL database with the exception of k-mers, which are stored using Elasticsearch. Users can access these results graphically through a web front end (staphopia.emory.edu) or programmatically through a web API. We have also written a R package (staphopia-R) to access the API. The pipeline has also been encapsulated into a Docker image, simplifying installation and running on local machines and in the cloud. More information about Staphopia is available at staphopia.emory.edu. All code is available on public GitHub repos.

#5: Toxicant Exposomics of a Key Bacterial Gut Commensal Carrier of Antibiotic Resistance

Presenting Author: **Stephen P. LaVoie**
Email: slavoie5@uga.edu

Complete Author List: Stephen P. LaVoie (1); Andrew G. Wiggins (1); Anne O. Summers (1)

(1) Department of Microbiology, University of Georgia, Athens, GA 30602, USA

Abstract:

Multi-antibiotic resistant (MAR) bacteria cost billions in medical care and many thousands of lives annually worldwide. Perennial calls to curtail agricultural antibiotic use and to fund new antibiotic discovery have yet to stem the resistance tsunami. That anything other than antibiotic use could drive the spread of MAR bacteria is heresy. However, it has long been known that exposure to common non-antibiotic toxicants promotes evolution of linked arrays of resistances to antibiotics and other toxic substances. The best studied example of a non-antibiotic toxicant enriching MAR is the toxic heavy metal mercury (Hg) to which most of the US population is exposed frequently via seafood and/or continuously via dental amalgam fillings. Widely found transposases and integrases assemble linked arrays of Hg and antibiotic resistance genes on mobile plasmids. Many animal and human lab studies and recent genomic epidemiology firmly support that widespread exposure to toxic metals co-selects mobile MAR in commensal and pathogenic bacteria. We studied metabolomic, proteomic, and transcriptomic effects of mercurials on the common gut commensal, E. coli, an opportunistic multi-resistant cause of urogenital and systemic infections. Specifically, we (1) quantified disruption of electrolyte, metal, and biothiol homeostases by acute inorganic Hg and organic Hg (RHg) compounds; (2) devised a novel global LC/MS-MS proteomics method to identify >300 proteins vulnerable to stable modification by acute RHg or Hg; and (3) used RNA-seq to discover strikingly different global gene expression after sub-acute exposure to RHg or Hg compounds. A notable finding was up-regulation of chromosomally encoded multi-antibiotic resistance genes in a plasmid-free strain of this gut bacterium, “priming” it, like other stressors, to withstand subsequent antibiotic exposure. Building on this foundation we are currently doing Hg-exposure transcriptomics of the same E. coli strain carrying the 100 kb conjugative plasmid NR1. This plasmid encodes the Hg resistance (mer) operon, which is genetically linked to the multi-antibiotic resistance IntI1 integron in the 19 kb transposon, Tn21. Our findings will inform pharmaceutical and nutritional interventions to prevent or minimize Hg-provoked co-selection of multi-antibiotic resistant bacteria in the human GI tract.

#6: A diverse microbial consortium drives nitrogen loss in large-scale aquarium sulfur reactors

Presenting Author: **Andrew S. Burns**
Email: andrew.burns@biology.gatech.edu

Complete Author List: Andrew S. Burns (1); Cory C. Padilla (1); Zoe A. Pratte (1); Kailen Gilde (2); Matthew Regensburger (2); Eric Hall (2); Alistair D.M. Dove (2); Frank J. Stewart (1)

(1) School of Biological Sciences, Georgia Institute of Technology
(2) Georgia Aquarium

Abstract:

High levels of nitrate, resulting from the decomposition of organic material, can have a deleterious effect on the health of inhabitants of aquarium systems. While nitrate is readily produced by aerobic nitrifying prokaryotes in aquariums, removal of nitrate is a more difficult process. Typically, nitrate is removed by removing and replacing a portion of the water mass. Water removal and replacement for large aquarium systems such as the Ocean Voyager exhibit at the Georgia aquarium (23,814,000 liters) is not feasible. In these systems, water is shunted through anaerobic vessels containing sulfur and aragonite where denitrifying bacteria facilitate nitrate removal. These bacteria use reduced sulfur species—such as sulfide, sulfite, and thiosulfate—as electron donors for the reduction of nitrate to dinitrogen gas through the intermediates nitrite, nitric oxide, and nitrous oxide. For the Ocean Voyager exhibit, two independent pads—each containing four anaerobic vessels—are used to process 491,400 liters of exhibit water per minute. A multi-faceted 'omics' approach was implemented to determine the microbial, gene, and transcript diversity across multiple potential niche spaces. The diversity of microbial species inhabiting the sulfur pellets, aragonite pellets, and interstitial water for each tower was determined using 16S rRNA amplicon sequencing. Metatranscriptomic and metagenomic sequencing was further applied to two towers from each pad. Distinct communities were formed in each of the independent pads. In one pad, a single operational taxonomic unit (OTU) closely related to Thiobacillus was a major component of the community while Thiobacillus was only present in low levels in the second pad. A major component of both pads was a consortium of numerous Sulfurimonas OTUs with no single Sulfurimonas OTU being dominant in either pad. Differences in the different physical niches were also evident in the denitrification systems. Sulfurimonas OTUs showed highest representation in interstitial water samples compared to the sulfur or aragonite pellets. Metagenomic and metatranscriptomic sequencing also showed diverse gene function both between microbial members of the community and specific niche space. The availability of multiple niches within the physical structure of the denitrification towers as well as the diverse metabolic potential of sulfur oxidation and denitrification created a diverse community facilitating the removal of nitrate. Overall, the use of multiple approaches allowed for the determination of both the microbial and genetic diversity within these complex systems.

#7: Community assembly in reef fish gill microbiomes

Presenting Author: **Zoe A. Pratte**
Email: zoe.pratte@biology.gatech.edu

Complete Author List: Pratte ZA (1); Hollman RD (1); Besson M (2); Stewart FJ (1)

(1) School of Biological Sciences, Georgia Institute of Technology
(2) Le Centre de Recherches Insulaires et Observatoire de l'Environnement de Polynésie Française

Abstract:

Teleost fish represent the largest and most diverse of the vertebrate groups and play important roles in food webs, as ecosystem engineers, and potentially as reservoirs for microbial communities (microbiomes). While the intestinal microbiome of fish is becoming recognized as a unique microbial niche with potentially key effects on host health, the microbiome of other body sites remains largely unexplored. Notably, while the microbiome of the gill is likely distinct from that of the water and potentially host-specific, the full diversity of the gill niche and the factors structuring this diversity remain unknown. Here, we provide the first comprehensive analysis of the fish gill microbiome. We focus on fishes common to coral reefs, sampling the gills and intestines of 207 adult fish and 55 juvenile recruits across 15 families and 53 species from Moorea, French Polynesia. Gill microbiome composition was significantly different from that of the gut in both adults and juveniles, with fish-associated niches showing lower alpha diversity and higher dispersion compared to seawater, sediment, and algae-associated microbiomes of the same habitat. Twenty-nine percent of microbial OTUs were only detected in fish samples, with 11% and 13% being specific to the gill or gut only, whereas 54% were only detected in the environment. Analysis of gill and intestinal microbiomes of the same individual, compared to between individuals, provided evidence of microbiome sharing among body site niches. These results identify key microbial taxa specific to the gill niche, connectivity between gill and gut niches, and evidence that the gill microbiome, like that of the gut, is shaped by host-specific organizing factors.

#8: The Microbiome of the Georgia Aquarium Ocean Voyager Exhibit

Presenting Author: **Nastassia V. Patin**
Email: npatin3@gatech.edu

Complete Author List: Zoe A. Pratte (1) Matthew Regensburger (2) Eric Hall (2) Kailen Gilde (2) Alistair D. M. Dove (2) Frank J. Stewart (1)

(1) School of Biological Sciences, Georgia Institute of Technology, Atlanta GA
(2) Georgia Aquarium, 225 Baker St NW, Atlanta GA

Abstract:

Marine microbes play critical roles in ecosystem health and function, including cycling of carbon, nitrogen, and other elements like sulfur. They also affect the health of fishes and other higher organisms that are in constant contact with the water column. Saltwater aquarium systems are designed to mimic natural marine environments as closely as possible, including housing resident microbes that process nitrogenous waste and prevent pathogenic infections. The Georgia Aquarium Ocean Voyager (OV) exhibit is a semi-controlled environment that mimics the basic physical and chemical parameters of the open ocean. It contains approximately 6.5 million gallons of artificial seawater with nutrient levels similar to those of an oligotrophic ocean basin. However, several important features differentiate the water column from a natural environment, including extensive filtration, sulfur-based denitrification, and ozone treatment. The habitat also contains large marine animals, including fish, sharks, and manta rays, but very few invertebrates or algal species. We present a 14-month time series of the OV microbiome and show that it is subject to bloom events featuring two heterotrophic bacterial taxa that are present only at very low levels in ocean environments. The relative abundances of these taxa are inversely correlated, suggesting they occupy a similar ecological niche in the water column. We reconstructed the genomes of these strains and highlight some interesting physiological characteristics. Notably, both bloom species contain genes for cyanophycin, an unusual amino acid polymer found frequently in cyanobacteria but rarely in heterotrophic bacteria. Cyanophycin is thought to act as a nitrogen and carbon storage compound, and the OV microbiome members have genes for both its synthesis and degradation. The genomic bins also feature plasmids that may facilitate the ecological plasticity of the dominant taxa. We also compared the OV microbiome with that of ocean environments with similar chemical features and found they differ dramatically, reflecting the inherently artificial nature of aquarium habitats. These findings have implications for large-scale aquaculture and demonstrate the opportunities for studying microbial ecology and evolution in a closed aquatic system.

#9: Community-wide transcriptional patterns reveal microbial endurance in a thermophilic composting process from Sao Paulo Zoo

Presenting Author: **Lucas P. P. Braga**
Email: lppbraga@iq.usp.br

Complete Author List: Lucas P. P. Braga (1); Rueben A Scriven (1); Andrew Thomas (1) Ronaldo B. Quaggio (1); Aline M. da Silva (1); João C. Setubal (1)

(1) Institute of Chemistry - University of Sao Paulo

Abstract:

Composting is a semi-engineered microbe-driven process for organic matter turnover. From a recent study performed by our group, we learned that microbial functions involved in biomass degradation shifts orchestrated in thermophilic and aerated composting chambers from Sao Paulo Zoo. However, biomass degrading functions are quite expected in such environments, and other underlying functions have been little explored so far. We hypothesized that a thermophilic composting process is performed by a microbiome under heat and oxidative stress conditions. To test this hypothesis, we analysed the metatranscriptomic time-series dataset obtained from the composting microbiome of the Sao Paulo Zoo facility. We then observed that microbial activity related with housekeeping functions and stress responses involved in protection against heat and oxidative damage, were dominant throughout the entire process. Interestingly, transposases and phage-host activity were also detected within the most active functions. The abundant transcripts related with heat-stress response were mainly of genes responsible for the chaperones CspA, GroES/GroEL, DnaJ/DnaK, and other heat-shock proteins (HSP20, HSP33, and HSP90). The transcripts of genes that could be directly associated with oxidative stress response, were genes transcribing catalases, peroxiredoxins, superoxide dismutases, and NADPH-dependent quinone oxidoreductase. In addition, higher proportions of integration host factor (ihfA, ihfB, and ihfC) activity, suggests that lysogenic cycle might be of considerable relevance for the composting microbiome. Network analysis revealed that shifts in functional activity during the process can be significantly (q-value<0.01) associated with shifts in taxonomic composition. However, when using a taxonomic classification with less resolution, it was observed that the nodes representing microbial clades were more clustered with the nodes representing gene transcripts; suggesting functional redundancy in the composting microbiome. So that taxonomic fluctuation at the OTU level is higher than the fluctuation at class or phylum level. Overall, the results obtained support microbial endurance in a thermophilic composting environment.

#10: The old N50, the new U50: pitfalls and remedies for bioinformatics analyses of complex human viromes

Presenting Author: **Terry Fei Fan Ng**
Email: ylz9@cdc.gov

Complete Author List: Terry Fei Fan Ng (1); Christina J. Castro (1,2); Rachel Marine (1); Roman L. Tatusov (3,4); Edward Ramos (3,4); Anna Montmayeur (3,4); W. Allan Nix (1); M. Steven Oberste (1)

(1) Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
(2) Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA
(3) NCIRD Core Bioinformatics Support, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
(4) CSRA International, Inc., Falls Church, VA, USA

Abstract:

Analyzing human clinical specimens with viral co-infections remains an analytical challenge in the field of bioinformatics. Robust next-generation sequencing (NGS) techniques provide an opportunity to investigate the different viral components of the co-infection. A crucial step for full viral genome analysis is de novo assembly. Analysis of viral NGS data is challenging for current de novo assembly programs, relative to resolving closely related viral taxa. Available de novo assembly programs typically output either contig fragments for each taxa, or one large contig combining all taxa. As a result, NGS data mis-assemblies and virus mis-identifications are common with samples that contain multiple taxa, especially if there are related types, strains and/or species present. A second issue is the lack of a performance metric that accurately depicts how well the assembly performed on viral datasets. Currently, such performance is measured by a metric called N50, but this often produces skewed, inaccurate results when complex viral NGS data are analyzed. We developed a new performance metric called U50. By comparing simulated and real datasets using U50 and N50, our results showed U50 has the following advantages: (1) reducing erroneously large N50 values due to poor assemblies, (2) eliminating greatly overinflated N50 values, (3) eliminating diminished N50 values caused by an abundance of small contigs, and (4) allowing comparisons across different platforms or samples using UG50%. Here we present an iterative mapping approach to handle complex viral NGS data, and a new performance metric called U50 to provide a better assessment of assembly output.

#11: Identification of nitric oxide (norB/Z) and nitrous oxide (nosZ) reductase gene fragments in soil metagenomes

Presenting Author: **Robert W. Murdoch**
Email: rmurdoch@utk.edu

Complete Author List: Robert W. Murdoch (1), Huaihai Chen (2), Zamin K. Yang (2), Christopher W. Schadt (2), and Frank E. Löffler (1,3)

(1) Center for Environmental Biotechnology, University of Tennessee, Knoxville, TN, USA
(2) Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
(3) Department of Microbiology, University of Tennesee, Knoxville, TN, USA

Abstract:

Nitrous oxide (N2O) is a greenhouse gas with ozone destruction potential. Microbial processes (denitrification) produce N2O in soils, but its consumption is limited to a single process: the reduction to dinitrogen (N2) catalyzed by N2O reductase (NosZ). Rapid and episodic N2O release from soils, occurring on an order of hours/days and apparently linked to dry-wetting and freeze-thawing events, has been observed. This collaborative LDRD project between ORNL & UTK investigators aims to link the presence and expression of norB/Z and nosZ to N2O release events induced in controlled soil mesocosms and a managed field site in TN. To characterize the field sites at the onset of the study, Illumina-based sequencing of DNA from soil samples collected from replicate plots at two soil depths (0-5 cm and 5-15 cm) was performed. Current efforts focus on the recovery of fragments of genes of interest (i.e., norB/Z, nosZ) from the metagenomic information with the goal of designing sets of qPCR primers with varying degrees of phylogenetic specificity.

#12: Genomic characterization and prioritization of nitrogen-fixing bacteria biofertilizers isolated from Colombian sugarcane fields

Presenting Author: **Luz K. Medina Cordoba**
Email: luz.medina@gatech.edu

Complete Author List: Luz K. Medina (1), Aroon T. Chande (1,2), Lavanya Rishishwar (2), Leonard W. Mayer (2), Joel E. Kostka (1,4) and I. King Jordan (1,4)

(1) School of Biological Sciences, Georgia Institute of Technology Atlanta, GA
(2) Applied Bioinformatics Laboratory, Atlanta, GA
(3) PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia
(4) Schools of Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA.

Abstract:

Sugarcane (Saccharum spp.) are tall, perennial grasses cultivated in tropical and warm temperate regions in Colombia, South America. Sugarcane production and processing is a multi-billion dollar/year business, supporting food, energy, and other industries around the country. Previous studies have shown that sugarcane harbors diverse plant growth promoting microorganisms, including nitrogen-fixing bacteria, which have the potential to serve as biofertilizers. The use of biofertilizers in sugarcane agriculture is key to reducing dependence on environmentally damaging and expensive chemical fertilizers. We are collaborating with the Colombian sugar cane company INCAUCA to isolate and characterize native nitrogen-fixing bacteria, with the aim of deploying them as plant growth promoting biofertilizers, and we previously isolated 23 nitrogen-fixing bacteria from INCAUCA sugar cane fields. The goal of this study was to use whole genome sequence analysis of these isolates in order to prioritize them with respect to their potential as biofertilizers. To this end, we are looking for strains that are predicted to have maximum benefit to the plants while presenting minimum risk to the environment, including local human populations. Functional genome annotations were used to prioritize strains that are enriched for nitrogen fixing and other plant growth promoting genes and depleted for virulence factors and antibiotic resistance genes. Comparative whole genome analysis revealed that 15 of 23 isolates belong to the genus Klebsiella, and 5 of 23 belong to genera closely related to Klebsiella. Functional annotation showed that all 23 isolates encode transcriptionally active nif operons, which are required for nitrogen fixation. These genomes also encode a variety of phosphate solubilization and siderophore production operons, as well as other genes involved in the synthesis of plant growth promoting metabolites (acetoin and butanediol). The isolates also contain antibiotic resistance genes (ampicillin, levofloxacin, and meropenem) and other virulence factors (host attachment factors and endotoxin). We developed a quantitative scoring system to rank potential biofertilizers by their predicted and experimentally validated growth promoting phenotypes, potential environmental risks, and antibiotic resistance profiles. Future work will be done to experimentally validate predicted biofertilizer activity along with potential virulence and antibiotic resistance of these isolates.

#13: Detection of Shiga-toxin genes in human stool metagenome short reads

Presenting Author: **Sung B. Im**
Email: sim8@gatech.edu

Complete Author List: Sung B. Im (1,2); Aroon T. Chande (1,3); Heather A. Carleton (2); Lavanya Rishishwar (3), Irving King Jordan (1,3)

(1) Georgia Tech, Atlanta, GA.
(2) Centers for Disease Control and Prevention, Atlanta, GA.
(3) Applied Bioinformatics Laboratory (ABiL), Atlanta, GA.

Abstract:

Culture independent diagnostic tests (CIDTs) continue to be widely adopted in the clinical setting as a faster, cheaper alternative to diagnose patients afflicted by foodborne illness. Largely due to CIDTs, the surveillance capacity of PulseNet, a national network of laboratories that monitors for emerging foodborne outbreaks by collecting and characterizing bacterial isolates sourced from foodborne illness cases, is under immediate threat by the on-going reduction of bacterial isolations. Escherichia coli (E. coli) is a common gut bacteria that can become pathogenic to humans if it acquires the necessary virulence factors. The phage induced Shiga-toxin gene is a distinct virulence determinant commonly found in the chromosome of enterotoxigenic Shiga-toxin producing E. coli (STEC). Identification of the Shiga-toxin gene variant is useful for foodborne outbreak investigations and surveillance. This study seeks to explore the potential associated with extracting Shiga toxin gene variant directly from metagenome sequence reads of disease state human stool sampled from patients infected with STEC. Allele identification was performed using the software package, STing, a k-mer based allele detection tool that works on minimally processed, raw sequence data. A database of Shiga-toxin gene sequence variants, developed at the Center for Genomic Epidemiology was used as the reference database. Allele identification was initially performed on raw sequence data from isolated E. coli. 1000 pathogenic E. coli strains sourced from diseased patients were selected. The strains were sequenced and made publically available by the Enteric Diseases Laboratory Branch and participating PulseNet affiliated laboratories from 2012-2016. The sample set includes diarrheagenic E. coli, determined by PCR to carry (n=925) or be without (n=75) Shiga-toxin genes. Allele identification was then performed on simulated metagenome sequence reads spiked with each of the 1000 STEC strains to investigate the allele identification capabilities of STing in mixed culture sequence reads. Metagenome data sets to be analyzed include patient stool samples from the 2011 Shiga-toxin producing E. coli O104 outbreak that took place in Germany, environmental samples from spinach spiked with outbreak associated STEC strains, and patient samples from the 2013 Colorado-Alabama Salmonella Heidelberg outbreak. Allele identification is determined by STing, if reads map to ≥95% of the predicted allele. STing identified the correct Shiga-toxin gene variant with >90% accuracy in the raw isolate sequence reads. As well, allele identification in the simulated mixed-culture sequence reads was in direct concordance to the known Shiga-toxin gene variant of the spiked isolate strain. Investigation of Shiga toxin allele identification in disease state stool metagenome sequence reads are currently on-going at the time of this writing.

#14: Ultrafast sequence typing with in silico DNA aptamers

Presenting Author: **Hector F. Espitia-Navarro**
Email: hspitia@gatech.edu

Complete Author List: Hector F. Espitia-Navarro (1,2); Aroon T. Chande (1,2,3); Heather Smith (4); King Jordan (1,2,3); Lavanya Rishishwar (1,2,3)

(1) School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
(2) PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia
(3) Applied Bioinformatics Laboratory, Atlanta, GA, USA
(4) School of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA.

Abstract:

Next-generation sequencing (NGS) has brought new possibilities for epidemiology such as identification of sequence type (ST) of bacterial pathogens and its characterization from a single data set. However, before an isolate's ST can be determined or genes of interest can be detected, NGS data require quality control, genome assembly and sequence similarity searching. These are computationally intensive and time-consuming steps, which are not ideally suited for real-time molecular epidemiology. We recently developed stringMLST, an assembly- and alignment-free, lightweight, platform-independent program capable of rapidly typing bacterial isolates. While stringMLST was more accurate and order of magnitude faster than its contemporary genome-based Multi Locus Sequence Typing (MLST) detection tools, it had limited resolution regarding the number of loci and scaled inefficiently for higher order locus-based typing schemes. To address these concerns, we substantially improved the stringMLST algorithm and created a new ultrafast typing and gene detection tool: STing. STing depends on exact pattern matching of substrings (k-mers) using Enhanced Suffix Arrays (ESA) indices as databases. After multiple searches of DNA k-mers in the ESA index, STing determines the isolate's ST by finding the allele for each locus of a typing scheme with the maximum number of k-mer hits. For detection, STing predicts presence/absence of genes by selecting the sequences with k-mer hits and a minimum length coverage of 75%. We used Illumina reads samples of *C. jejuni* (n=10) and *N. meningitidis* (n=999) to test the typing accuracy of STing using the MLST scheme (loci=7). STing correctly predicted the STs in 100% of the samples for each species. We also tested our tool in two larger schemes, rMLST and cgMLST (loci=53 and 1,605), using 20 Illumina reads samples of *N. Meningitidis* in each of them. STing correctly predicted 97.1% and 92.6% of the alleles for rMLST and cgMLST, respectively. Depending on the scheme, STing was 4.7x to 50.8x faster and consumed 4.5x to 14.7x less RAM than stringMLST. To test the ability to detect genes, we selected a set of AMR genes (n=16) present on 12 bacterial genomes from which we generated positive and negative read samples at 20x and 40x sequencing depth. STing correctly identified AMR genes in the positive samples and predicted the absence of any AMR gene in the negative samples. The STing algorithm scales efficiently to genome-scale typing schemes, and it is both more accurate and much faster than its predecessor. STing provides additional utility for fast gene detection directly from NGS read sequences, with applications in culture-free diagnostics as well as virulence factor and antimicrobial resistance profiling.

#15: Symbiont taxonomic and genetic diversity in lucinid (Bivalvia:Lucinidae) species

Presenting Author: **Jean Lim**
Email: jslim@g.clemson.edu

Complete Author List: Jean Lim (1); Annette Engel (2); Laurie C. Anderson (3); Barbara J. Campbell (4)

(1) Department of Biological Sciences, Clemson University, Clemson, SC
(2) Department of Earth and Planetary Sciences, University of Tennessee Knoxville, Knoxville, TN
(3) Department of Geology and Geological Engineering, South Dakota School of Mines & Technology, Rapid City, SD
(4) Department of Biological Sciences, Clemson University, Clemson, SC

Abstract:

Bivalve species in the family Lucinidae harbor gill-associated symbionts belonging to the class Gammaproteobacteria, which are capable of sulfur oxidation and carbon fixation as thioautotrophs. Co-symbiosis in lucinid clams has not been comprehensively tested. The genetic repertoire of lucinid symbionts is also only beginning to be investigated. Using 16S rRNA gene and metagenomic sequencing, we analyzed the microbiomes and metagenomes of three lucinid species: Ctena orbiculata, Stewartia floridana, and Phacoides pectinatus, collected from various sites in Florida, USA. At the 16S rRNA gene level, in addition to retrieving genes related to the thioautotrophic symbiont species, we observed the consistent presence of taxa enriched in the gill environments that could represent potential symbionts. Of these taxa, >80% completed genomes of two previously unsequenced species from the order Oceanospirillales and Spirochaetales were assembled from the gill metagenomes of P. pectinatus, with read coverage abundances consistent with their relative abundances in the 16S rRNA gene dataset. We also assembled >90% complete draft genomes of four previously unsequenced thioautotrophic symbiont species from the lucinid gill metagenomes, including two distinct species co-existing in the C. orbiculata population. Interspecific genetic diversity among the symbiont species was observed for nitrogen-, carbon-, and sulfur-related and a shared core set of thioautotrophic, heterotrophic, and respiratory genes. Genomic analyses revealed genes for diazotrophy and previously unreported methylvory in the thioautotrophic symbionts of C. orbiculata and S. floridana. Overall, findings from this study suggest that taxonomic and genetic diversity among lucinid symbionts are higher than previously thought. Our results warrant further investigations into co-symbiosis in lucinids, as well as symbiont genetic diversity within a local host population and across geographically segregated host populations.

#16: Genomic and functional diversity of SAR11 in the Delaware Bay

Presenting Author: **Barbara J. Campbell**
Email: bcampb7@clemson.edu

Complete Author List: Barbara J. Campbell (1); Jean S. Lim (1); David Kirchman (2)

(1) Department of Biological Sciences, Clemson University, Clemson, SC
(2) School of Marine Science and Policy, University of Delaware, Lewes, DE

Abstract:

SAR11 are a ubiquitous, extremely abundant and diverse clade of marine planktonic bacteria. Their streamlined genomes and lack of many regulatory systems suggest they may not respond to changing environmental conditions. Here we assembled six distinct SAR-11 affiliated genomes from metagenomic samples collected from surface water along the Delaware Bay in different seasons, salinities, times of day, and size fractions. We then mapped metatranscriptomic reads from this variety of environmental conditions to these metagenome assembled genomes (MAGs). The estimated genome completeness of the six SAR11 MAGs ranged from 78 to 95% and all were less than 1.8% contaminated via CheckM. Using a phlyogenomics approach, we determined that these six MAGs belong to the SAR11 subclades Ia, II, IIIa and V. All of the SAR11 MAGs contain genes/pathways to potentially oxidize or produce volatile organic compounds. Two SAR11 MAGs, one from 1a and one from IIIa, contain genes involved in CO oxidation. Surprisingly, we found many differences in SAR11 gene expression via RNAseq analysis. The largest differences in individual MAG gene expression occurred between spring and summer, where up to 73% of expressed genes were significantly different. Many of these were transporters for sugars, ammonium or phosphorous, or were involved in growth, including cell wall production, cell division and protein synthesis. Genes involved in DMSP breakdown and carbon monoxide utilization were also differentially expressed between spring and summer in some of the SAR11 MAGs. Our work demonstrates that SAR11 found in the Delaware Bay are very diverse in their phylogeny and genetic potential and indicates more functional gene regulation in situ than previously recognized in cultured organisms.

#17: Predicting Disturbance-driven Impacts on Ecosystem Services in Coastal Wetlands

Presenting Author: **Suja Rajan**
Email: srajan@ua.edu

Complete Author List: Suja Rajan(1), Patrice Crawford(1), Alice Kleinhuizen(1, 2), Behzad Mortazavi(1, 2), Patricia A. Sobecky(1)

(1) The University of Alabama, Department of Biological Sciences, Tuscaloosa, AL
(2) Dauphin Island Sea Lab, Dauphin Island, AL

Abstract:

Natural and human-induced disturbances pose significant threats to the health and long-term productivity of Alabama coastal wetlands. As wetlands are a vital state resource, decisions on management, restoration, and remediation require actionable data if socio-economic demands are to be balanced with efforts to sustain these habitats. In 2010, the BP oil spill was a large and severe disturbance that threatened coastal Gulf ecosystem services. The largest marine oil spill to date served to highlight fundamental gaps in our knowledge of oil-induced disturbances and the resiliency and restoration of coastal Alabama wetland functions. To address these gaps, a year-long mesocosm study was conducted to investigate oil-induced effects on (i) plant-microbial interactions, (ii) microbial and plant biodiversity, and, (iii) the contributions of microbial genetic biodiversity to ecosystems services. In this study, Avicennia germinans (black mangrove), a C3 plant that grows from the tropics to warm temperate latitudes, were grown with or without mono- and polyculture mixtures of Spartina alterniflora, a C4 plant. At an interval of 3-months, 1.9 L m-2 of Louisiana sweet crude oil was introduced as a pulse disturbance. Molecular based analyses of microbial community biodiversity, genetic diversity, and functional metabolic genes were compared to controls (i.e., no oil disturbance). To assess the oil-induced effects on the nitrogen (N) cycle, measurements of denitrification and N fixation processes were conducted. Our results showed that community diversity and phylogenetic diversity significantly changed and that the oil disturbance contributed to the creation of niches for distinct microbial types. The abundance of N-fixing microbial types increased as the abundance of denitrifying microbial types decreased as a result of the oil disturbance. As denitrification is an ecosystem service that directly contributes to removing nitrate (NO3-) loading to coastal zones, impairment of this process is detrimental to the long-term health and productivity of the Gulf of Mexico. Our results are designed to investigate controlling factors and yield insights to aid decision-makers in their ongoing management efforts to restore wetlands along the Alabama coast and elsewhere.

#18: An emerging human fungal pathogen, multi-drug resistant Candida auris

Presenting Author: **Xin Huang**
Email: xin.huang@nih.gov

Complete Author List: Xin Huang (1); Rory Welsh (2); Meghan Bentz (2); Rebecca Drummond (3); Sean Conlan (1); Clayton Deming (1); Weng-lan Ng (1); Michail Lionakis (3); Anastasia Litvintseva (2); Julie Segre (1)

(1) National Human Genome Research Institute, NIH, Bethesda, MD
(2) National Center for Emerging and Zoonotic Infectious Diseases, CDC
(3) National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD

Abstract:

First identified only eight years ago as disperse clinical cases in Japan, Venezuela, South Africa and India, Candida auris has emerged as an urgent threat to patients globally. Candida auris is a human fungal pathogen, which is resistant to multiple classes of antifungal agents, able to spread in a hospital settings and causes fatal bloodstream infections. Preliminary studies suggest that skin is the major site of colonization, and that patients remain colonized for months. From the skin, C. auris is shed into the environment, where it persists for days and possibly weeks. All of these factors prompted both Public Health England and the US Center for Disease Control and Prevention (CDC) to issue multiple clinical alerts in 2016 and 2017. We aim to establish a C. auris translational research program that analyzes data from C. auris colonized patients and then establishes mouse models to assess risk factors for colonization and infection. Our long-term goal is to utilize the mouse model to test new methods and possibly new drugs to decolonize and treat patients. Working in collaboration with the CDC, we performed microbial sequencing and analysis of patient skin samples from two long term care facilities, currently experiencing C. auris outbreaks. To characterize the fungal and bacterial communities, we used amplicon sequencing of the fungal internal transcribed spacer 1 (ITS1) and bacterial 16s rRNA regions. Based on reference genomes, we established the bioinformatics pipeline to identify Candida ITS1 sequences to the species level, providing clear resolution of C. albicans, C. auris, C. glabrata, C. parapsilosis, etc. Strikingly, from the healthcare outbreak skin samples, Candida species represented greater than 60% of the sequence reads for 11 of 20 patients, including patients who were positive or negative for C. auris in both facilities. This is in contrast to our survey of fungal diversity amongst healthy volunteers in which we found extremely low (<0.01%) colonization with Candida spp. and no C. auris. Bacterial 16s amplicon sequencing also revealed dysbiosis and colonization with species associated with healthcare-associated pathogens. We have performed a pilot study with wild-type mice, in which we colonized intact skin with C.auris based on a protocol developed for microbial topical association. Long-term colonization of the mice was assessed with both culturing and fungal ITS1 sequencing of skin swabs. Our next mouse experiments will test risk factors that have been potentially identified in human patients, including underlying conditions, receipt of antimicrobials and immune status.

#19: Discovery of novel viral families in the twilight zone of protein sequence similarity

Presenting Author: **Natalya Yutin**
Email: yutinn@ncbi.nlm.nih.gov

Complete Author List: Natalya Yutin (1) and Eugene V. Koonin (1)

(1) National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract:

Viruses are the numerically dominant biological entities in the biosphere. Viral genomes are relatively small, so that assembly of a complete viral genome from metagenomic sequences is often feasible. Viruses typically evolve much faster than cellular life forms, which makes identification of homologs and functional annotation of viral proteins a challenging task. Despite the huge amount of sequenced genomes and metagenomes present in publicly available databases, many viral proteins are reported as having no homologs, and many viral genomes are not recognized as such. By deep searches of GenBank databases (both nr and wgs), we connect distant homologs of viral hallmark proteins, which allow us to identify new viruses, annotate viral genomes, and trace their evolutionary relationships. Here we present two recent cases: the discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut, and the diversity of viral genomes encoding Tectivirus-like major capsid proteins throughout the prokaryotic world.

#20: Genomes Galore – Core Techniques, Libraries, and Domain Specific Languages for High-Throughput DNA Sequencing

Presenting Author: **Tony C. Pan**
Email: tcp1975@gmail.com

Complete Author List: Tony C. Pan (1), Patrick Flick (1), Chirag Jain (1), Nagakishore Jammula (1), Sriram P. Chockalingam (1), Sharma V. Thankachan (2), Srinivas Aluru (1)

(1) Georgia Institute of Technology, School of Computational Science and Engineering
(2) University of Central Florida, Department of Computer Science

Abstract:

Advances in next generation sequencing technologies have dramatically reduced the cost and improved latency and throughput. However, few bioinformatics tools can efficiently process the datasets at the current generation rate of 1.8 terabases per 3-day experiment from a single sequencer. Increasing adoption of genomic data analysis in diverse disciplines including biology, agriculture, personalized medicine, and environmental sciences, further requires efficient big-data management and analytic capabilities that are afforded by distributed memory and high core count systems. With the support of NSF Big Data initiative, we have developed a suite of algorithms, implementations, and tools, under the project ParBLiSS, to support large scale genomic data analysis in distributed parallel environments. We have developed algorithms for indexing and querying biological sequences that have applications in genome alignment and mapping, genome assembly, and error correction. For indexing and counting fixed length subsequences called k-mers, the Kmerind k-mer indexing library outperforms all existing tools in both shared memory and distributed memory environments. For dynamic indexing and sequence matching, the PSAC library constructs a suffix array more than 100 times faster than the fastest sequential algorithm, using 1024 cores. Our algorithm for k-mismatch all-pairs maximal common substring, useful for alignment, clustering and assembly applications, represents the first efficient parallel algorithm for such problem. We have also developed an efficient parallel algorithm for partitioning de Bruijn graphs to enable parallelization of metagenome analysis and assembly. While our parallel algorithms have been developed for distributed memory environments, they have also been shown to perform well on shared memory systems. Our libraries are have been developed to be high performance, flexible, and extensible. They are available at https://github.com/ParBLiSS under Apache open source license.

#21: High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries

Presenting Author: **Chirag Jain**
Email: cjain7@gatech.edu

Complete Author List: Chirag Jain (1,4); Luis M. Rodriguez-R (2,3); Adam M. Phillippy (4); Konstantinos T. Konstantinidis (2,3); Srinivas Aluru (1,5)

(1) School of Computational Science and Engineering, Georgia Tech, Atlanta GA
(2) School of Civil and Environmental Engineering, Georgia Tech, Atlanta GA
(3) School of Biological Sciences, Georgia Tech, Atlanta GA
(4) National Human Genome Research Institute, NIH, Bethesda
(5) Institute for Data Engineering and Science, Georgia Tech, Atlanta GA

Abstract:

Rapid developments in high-throughput sequencing technologies have led to an exponential increase in the number of available microbial genomes and metagenomes. New computational algorithms that better scale with the available genomics data are highly needed. This work presents a novel scalable algorithm for an important problem in microbiology: computing the whole-genome genetic relatedness among genomes. Whole-genome Average Nucleotide Identity (ANI) is a widely trusted metric for assessing the relatedness among microbial genomes and delineating species as it yields more robust results than traditional methods used for the same purposes (e.g., DNA-DNA hybridization and 16S rRNA gene sequencing). However, current alignment-based ANI tools are computationally expensive for comparisons of thousands of genomes. We present a novel and faster method, FastANI, to compute ANI using alignment-free approximate sequence mapping. Our benchmarks showed that FastANI produces an accurate ANI estimate, and is up to three orders of magnitude faster than the BLAST-based approach. We applied FastANI to answer a fundamental question in prokaryotic evolution: Is there a continuum of genetic diversity among prokaryotic genomes or clear species boundaries prevail instead? We computed pairwise ANI values among all prokaryotic genomes available in NCBI database in less than one week running time on a medium-sized computation cluster. Our results revealed a clear genetic discontinuity among database genomes around 85-95% ANI, i.e., <0.2% of the total 8 billion genome pairs showed between 85-95% ANI, revealing that species boundaries may exist among prokaryotic genomes.

#22: Spatial and temporal characterization of river sediment microbial communities using time-series metagenomics

Presenting Author: **Brittany Suttner**
Email: bsuttner3@gatech.edu

Complete Author List: Brittany Suttner (1); Janet K. Hatt (1); Michelle Carter (2); Micheal Cooley (2); Konstantinos T. Konstantinidis (1)

(1) Georgia Institute of Technology, Atlanta, GA, USA
(2) USDA-Agricultural Research Service, Albany, CA, USA

Abstract:

The Salinas River valley is one of the most productive agricultural regions in the US. Hence the quality of the water flowing through this region is regularly monitored to ensure its safety for agricultural applications as many cases of O157:H7 outbreaks have been associated with produce contaminated by irrigation water. Here we describe our efforts to draw correlations between detection of pathogens by traditional culture based methods and environmental factors (e.g. rainfall, season, etc.) and shotgun metagenomes, and assess if anthropogenic impacts are more prevalent in the agriculturally-impacted sediment metagenomes compared to more pristine locations. For this, the water/sediment interphases from three sites along a creek in the Salinas River Valley have been selected for a time series metagenomics study. Two of the sites are impacted by cattle ranching and agriculture, while one upstream site has no apparent anthropogenic influences. Overall, our results suggest that these river sediment communities harbor comparable, if not greater, taxonomic and functional diversity than most soils. Although the effect of temporal variation was not significant, spatial separation was the strongest driver of diversity and community structure in this system. We were unable to detect any effect of environmental variables on the communities nor detect pathogens in the metagenomes of samples that had tested positive for pathogens by culture-based testing due to under-sampling of these highly diverse communities and despite sequencing about 5Gbp per sample with an Illumina platform. The estimated limit of detection of an E. coli genome in the metagenome samples was 0.6% per million reads using ImGLAD (Castro et al., in review), a new tool that predicts the probability that a target organism is present in a metagenome. Notably, a high background level of reads annotated as antibiotic resistance genes were found in all samples based on blast-homology searches, and appeared to be enriched in the upstream pristine samples, which suggests that these communities may be natural reservoirs for antibiotic resistance in the environment. In order to test for potential false positive matches to highly conserved regions of the reference antibiotic resistance genes, we used ROCker (Orellana; Nucl. Acid Res. 2017), a gene annotation tool that identifies position-specific thresholds for higher confidence alignments of short reads to reference genes. Our work showed that the blast-based approach greatly over-estimated the abundance of these genes by about 10 fold and that tetM, which encodes resistance to tetracyclines that are commonly used in livestock, was more prevalent in the upstream, pristine sites.

#23: Metagenome-guided isolation of Macondimonas diazotrophicus: A novel keystone species of oil biodegradation in beach sands affected by Macondo oil

Presenting Author: **Smruthi Karthikeyan**
Email: smruthi98@gatech.edu

Complete Author List: Smruthi Karthikeyan (1); Patrick Heritier-Robbins (1); Minjae Kim (1); Luis M Rodriguez-R (1,2); Janet K. Hatt (1); Will A. Overholt (2); Joel E. Kostka (2,3); Markus Huettel (4); Konstantinos T. Konstantinidis (1,2)

(1) Department of Civil and Environmental Engineering, Georgia Tech, Atlanta GA
(2) School of Biological Sciences, Georgia Tech, Atlanta, Georgia, USA
(3) School of Earth and Atmospheric Sciences, Georgia Tech, Atlanta GA
(4) Department of Earth and Atmospheric Sciences, Florida State, Tallahassee, FL

Abstract:

The Deepwater Horizon oil spill released millions of barrels of oil and massive amounts of natural gas to the Gulf of Mexico which had a profound impact on the indigenous microbial communities in its vicinity thereby shaping it to respond to these oil induced perturbations. Our previous analysis of 16S rRNA gene amplicon data and the functional gene content of time series metagenomic data originating from oiled beachsands in Pensacola Municipal Beach, FL (Rodriguez-R et al., ISME 2015) revealed a high abundance of nitrogen fixing genes (namely nifH) and the predominance of a few specific nifH alleles, a potentially important finding since oil biodegradation is often nitrogen-limited. A particular allele of the nifH gene had a dramatically higher abundance in comparison with the rest. In order to identify the genetic context of this gene, targeted population reconstruction using manual assembly complemented with PCR-walking for linking the 16S gene yielded a nearly complete genome sequence (bin) that included the abundant nifH gene allele along with the complete rRNA operon. Functional annotation of this genome revealed genes for hydrocarbon degradation, methanotrophy, biosurfactant production, nutrient scavenging and other related mechanisms that could enhance growth in oil-contaminated environments. The relative abundance of the genome bin rapidly increased from below detection levels in the clean/pre spill beachsand samples from Pensacola Beach to 29% of the entire community gene pool in oiled sand samples, returning to undetectable levels in the recovered sediments. SSU rRNA gene sequences of this genome were also widespread almost exclusively in oiled/hydrocarbon contaminated sediments across the globe including sediments impacted by the DWH oil spill. Targeted efforts to isolate this organism from oiled Pensacola beachsands were successful, yielding a rod-shaped bacterium that showed 99.8% genome-aggregate average nucleotide identity (ANI) to the previously identified population bin. Whole genome comparisons to available genomes revealed that this genome represents a novel family within Gammaproteobacteria, for which we propose the name Macondimonas diazotrophicus. The ecological distribution and metabolic versatility of M. diazotrophicus coupled to its abundance patterns during oil biodegradation suggest that it may potentially represent a keystone species of oil biodegradation and a promising biomarker for oil contamination and ecosystem recovery.

#24: Widely used disinfectants can promote antibiotic resistance

Presenting Author: **Minjae Kim**
Email: minjaekim45@gmail.com

Complete Author List: Minjae Kim(1), Michael R. Weigand(1), Seungdae Oh(1), Janet K. Hatt(1), Spyros G. Pavlostathis(1), Konstantinos T. Konstantinidis(1)

(1) Georgia Institute of Technology, Atlanta, GA, USA

Abstract:

Whether disinfectant exposure promotes antibiotic resistance (AbR) has been a long debate with major practical consequences. To obtain insights into this issue, we exposed a microbial community originating from a contaminated river sediment (Calcasieu River, USA) to benzalkonium chlorides (BAC; a family of quaternary ammonium disinfectants) for 3 years in a fed-batch bioreactor receiving Dextrin Peptone plus BAC as the sole carbon source (DPB bioreactor). A bioreactor receiving only Dextrin Peptone (DP) served as a control. Further, Pseudomonas aeruginosa isolates, which originated from the same ancestral population in the original inoculum, were obtained from both bioreactors and used to study adaptive evolution in response to increasingly higher BAC concentrations. Metagenomics of the bioreactors and molecular cloning of resistance genes revealed that BAC exposure induced the spread of AbR in several species via horizontal transfer of mobile DNA elements that encode a BAC efflux pump together with AbR genes. Although several BAC-exposed isolates exhibited higher resistance to certain antibiotics, others did not, presumably due to their intrinsic resistance mechanisms. Genomics and transcriptomics analysis of P. aeruginosa strains revealed several fixed mutations in BAC-evolved populations such as in the histidine kinase A domain of the pmrB, which regulates resistance to polymyxin B, consistent with up to 8-fold higher MIC values for polymyxin B, and the overexpression of several AbR genes such as the MexCD-OprJ regulating tetracycline and ciprofloxacin resistance in response to BAC exposure. We will also report on the organisms that were isolated from our bioreactors and found to be robust BAC-degraders based on novel enzymatic pathways. Collectively, these results substantially advance the highly debatable issue of biocide-induced AbR and have implications on how to limit exposure to biocides and thus, the acquired antibiotic resistance.

#25: Understanding the underlying mechanisms of genome-wide selective sweeps in natural microbial populations using time-series metagenomic analysis.

Presenting Author: **Aditi S. Paranjpe**
Email: aditip@gatech.edu

Complete Author List: Aditi Paranjpe(1), Luis Rodriguez-R(2), Despina Tsementzi(2), Alexandra Meziti(2), Chengwei Luo(3), Konstantinos Konstantinidis (1,2)

(1) School of Biological Sciences, Georgia Tech, Atlanta, GA
(2) School of Civil and Environmental Engineering, Georgia Tech, Atlanta, GA
(3) Broad Institute of MIT & Harvard, Cambridge, MA

Abstract:

The identification of sequence-discrete natural microbial populations, which may correspond to genuine species, raises the important question of what might be the underlying driver(s) of population cohesion. Several theories have been advanced for explaining the origin and maintenance of sequence-discrete populations, including recombination frequency, ecological selective sweeps, population bottlenecks, and random birth/extinction. However, none of these theories have yet gained universal support mainly due to lack of experimental data. Time-series metagenomic studies can offer insights into the maintenance and dynamics of genotype diversity of sequence-discrete populations. For instance, genomic sequences of several cultured or uncultured representatives of a population can be queried against time-series community metagenomics data to identify, count, and follow over time the allele diversity variations. We analyzed a six-year-long time-series metagenomic dataset from Lake Lanier. From this dataset, a total of 140 microbial genomes having > 95% completeness and <5% contamination were recovered and queried against the available metagenomes in order to identify SNPs. Subsequently, the frequencies of the different alleles at each SNP position were quantified across the metagenomes to study the temporal dynamics of allele diversity. The analysis showed that most of the genomes (117/140) maintained the overall allele diversity over the years, with <10% change in allele diversity between any two time points. SNPs of only a few genes (<3% of the total genes) encoded by these genomes were gradually fixed, indicating possibly positive selection for the corresponding protein functions. Nonetheless, 23 microbial populations showed decrease in the allele diversity over time, indicating genome-wide selective sweeps. The SNP positions showing loss of allele diversity over time were present throughout the whole genomic sequence in non-coding as well as coding regions, indicating that these results are not due to gene sweeps (e.g., genes becoming fixed through horizontal gene transfer and recombination). Instead, these findings were consistent with strain replacement due to selective advantages (sweeps) or seasonal population bottlenecks. Collectively, our findings show that population sweeps are relatively uncommon within short periods of time in Lake Lanier (5-6 years), and advance current understanding of the mechanisms maintaining sequence-discrete populations.

#26: Detection of recent gene exchange among closely related bacterial genomes and implications for the bacterial species concept

Presenting Author: **Maria J. Soto-Giron**
Email: juliana.soto@gatech.edu

Complete Author List: Maria J. Soto-Giron(1); Juan C. Castro(1); Luis M. Rodriguez-R(2); Konstantinos T. Konstantinidis(2)

(1) School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA.
(2) School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA.

Abstract:

High-throughput sequencing has revealed that bacterial genomes are highly dynamic, driven mostly by horizontal gene transfer (HGT). Quantifying HGT and its role in bacterial genome evolution and speciation has been challenging, especially within species, due to the high sequence identity of core genes at this level (i.e., low signal-to-noise ratio). Further, phylogenetic-based methods for detecting HGT do not scale well with a large number of genomes (e.g., 100s) to analyze. Here, we devised a new method to estimate rates of recent HGT among closely related genomes based on the observed frequency of identical genes (F100) shared between two genomes relative to the number of such genes expected by chance according to their genome-aggregate average nucleotide identity (ANI) value. Results from comparisons of hundreds of available genomes showed that our approach can reliably estimate the genomic fraction under recent exchange between closely related genomes (ANI 95.00 to 99.97%). Particularly, the deviation of F100 from the average expected frequency can be applied to distinguish whether the population presents a recombinogenic, i.e., higher F100 values than the average as exemplified by ecologically versatile organisms, or clonal structure, i.e., lower F100 values than the average as exemplified by obligate endosymbionts. Moreover, our results indicate that on average, ~13% of the total genes in the genome have been recently exchanged in versatile organisms. Extremely high rates of recent genetic exchange in a few outlier genomes indicated that they might be undergoing primarily sexual speciation.

#27: Accurate typing of hundreds of *Bacillus anthracis* genomes using the Microbial Genome Atlas (MiGA) webserver

Presenting Author: **Konstantinos T Konstantinidis**
Email: kostas@ce.gatech.edu

Complete Author List: Luis M. Rodriguez-R (1,2); Angela Pena-Gonzalez (3), Chung K. Marston (4); Cari Beesle (4); Jay E. Gee (4); Alex Hoffmaster (4); Konstantinos T. Konstantinidis (1,2,3)

(1) School of Civil & Environmental Engineering, Georgia Tech, Atlanta, GA
(2) Center for Bioinformatics & Computational Genomics, Georgia Tech, Atlanta, GA
(3) School of Biological Sciences, Georgia Tech, Atlanta, GA
(4) Centers for Disease Control and Prevention, Atlanta, GA

Abstract:

The characterization, documentation, and exploration of prokaryotic genomic diversity within a single species have been recently enabled by the rapid progress in high-throughput sequencing technologies. However, the use of this data for determining phylogenetic structure and informing classifications at infra-specific levels remains a task that requires significant computational and human resources, only achievable with industrious efforts for a handful of microbial model species. Here, we present a methodology based on the concept of Average Nucleotide Identity (ANI) to leverage whole-genome comparisons in the estimation of phylogenetic structure, the proposal of uniform infra-specific classifications, and the rapid identification of novel genomes with high precision. We propose three techniques based on ANI for the exploration of infra-specific prokaryotic diversity. First, ANI clustering, consisting of the application of hierarchical agglomerative clustering, demonstrates the usefulness of 1-ANI as genome-wide distances and its tight correlation with phylogenetic distances. Second, ANI classification is an extension of ANI clustering for the detection of naturally forming clusters within the known diversity of a species, using k-medoid clustering and a selection criterion based on Silhouette widths. Finally, ANI typing is a method for the rapid traversing of ANI classifications based on distance-space search reduction that correlates closely with other typing methods but offers high precision, scalability, and stable classifications. We applied our method to a collection of 412 newly sequenced and 56 previously determined genomes from the clonal pathogen Bacillus anthracis. This group is particularly challenging due to its extremely low variation (ANI > 99.8% for all included genomes), hence representing a case in which whole-genome resolution is maximally challenged. Despite the challenges presented by this species, we observed a high correlation between 1-ANI and phylogenetic distances (Pearson's R = 0.92), and the ANI classification recapitulated other well-established clades based on multi-locus sequence analysis (MLSA) or curated sets of single nucleotide polymorphisms (SNPs). Finally, the best match of a novel query genome against the available reference database of 13,000 genomes can be obtained in about 30 minutes, on average, using ANI typing, while hundreds of query genomes can be processed in parallel. The introduced methodology was implemented as part of the Microbial Genomes Atlas (MiGA), available at www.microbial-genomes.org, and can be applied to any collection of genomes of the same species.

#28: Photosynthesis Genes in Viral Genomes Across a Freshwater-Saltwater Gradient in Southeast US

Presenting Author: **Carlos A. Ruiz-Perez**
Email: cruizperez3@gatech.edu

Complete Author List: Carlos A. Ruiz-Pérez(1); Despina Tsementzi(2); Janet K. Hatt(2); Matthew B. Sullivan(4); Konstantinos T. Konstantinidis(1,2,3)

(1) School of Biological Sciences, Georgia Tech, Atlanta, GA.
(2) School of Civil and Environmental Engineering, Georgia Tech, Atlanta, GA.
(3) Center for Bioinformatics and Computational Genomics, Georgia Tech, Atlanta, GA.
(4) Department of Microbiology, Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH.

Abstract:

Phages infecting cyanobacteria play an important role in controlling host population dynamics and modifying host genome. Some of these phages encode host-acquired functional genes, known as auxiliary metabolic genes (AMGs), that are thought to confer a fitness advantage because the expression of the corresponding proteins increases progeny yield during infection. AMGs involved in photosynthesis are widespread among marine phages, especially those belonging to the Myoviridae and Podoviridae families infecting Prochlorococcus and Synechococcus. Although photosynthesis-related AMGs have been extensively studied in oceans, their distribution, prevalence, diversity, and genomic context in freshwater habitats remain poorly understood. To advance understanding of these issues, we surveyed five freshwater lake systems and two estuarine locations in the Southeast US interconnected by the Chattahoochee River, for four years, using metagenomic sequencing of viral populations. Our results revealed the prevalence of photosynthesis-related AMGs in viral genomes recovered in all five lakes, albeit at relative abundances about 10 times lower compared to the ocean or estuarine locations, due presumably to different infection strategies and viral latent periods. Most of the genome fragments recovered belonged to Myoviridae and Podoviridae families. Myoviridae genomes showed the presence of a more diverse repertoire of photosynthesis AMGs compared to Podoviridae genomes, which only encoded the photosystem II D1 protein-encoding gene psbA and high-light inducible proteins. Our analysis also showed a limited exchange between marine and freshwater viral psbA sequences, while the estuarine locations appeared to harbor their own psbA sequence variants. Collectively, our findings indicated that the presence of AMGs involved in photosynthesis is a widespread strategy used by Myoviridae and Podoviridae cyanophages in aquatic ecosystems, and the role of the large understudied AMG diversity in viral ecological and infection strategies needs to be better understood.

#29: Viral community dynamics and an extensive virus-host interaction network revealed by a 5-year-long metagenomic time-series from a temperate freshwater ecosystem

Presenting Author: **Despina Tsementzi**
Email: despoina.tsementzi@gatech.edu

Complete Author List: Despina Tsementzi (1), Luis M Rodriguez R (1), Carlos Ruiz Perez (2), Dam Phuongan (3), Eberhard O Voit (3), Konstantinos T Konstantinidis (1,2)

(1) School of Civil and Environmental Engineering, Georgia Tech, Atlanta, GA
(2) School of Biological Sciences, Georgia Tech, Atlanta, GA
(3) Department of Biomedical Engineering, Georgia Tech, Atlanta, GA

Abstract:

Viruses play a pivotal role in the maintenance and genomic plasticity of microbial communities, with direct impacts on population control, horizontal gene transfer, community diversity and metabolism. While viral metagenomics has recently helped mapping large-scale spatial variability and sequence diversity, these efforts have largely been restricted to the oceans and only a snapshot of samples. Here we analyze 17 paired viral and microbial metagenomes from Lake Lanier (GA, USA) to map diversity and virus-host interaction dynamics over the span of five years. The viral α-diversity showed high correlation with that of the microbial community, increasing during fall and decreasing during the early summer. Despite the low portion of sequences with matches against public databases (~35% of assembled contigs), assembly and clustering of the time-series viromes enabled us to identify ~1,500 nearly-complete genome sequences representing individual viral populations. Most of these populations showed highly variable abundance profiles but long-term persistence. The taxonomic composition was more similar to other freshwater viromes than to marine or hypersaline viral communities, and the functional distribution was highly skewed in comparison to the microbial communities. Additionally, the deeply sequenced microbial genomes allowed the isolation of 1,126 nearly complete microbial genomes from metagenomic binning. Lotka-Voltera (LV) dynamic modeling of the bacterial and viral genomes identified a couple hundred viral-host links, some of which were supported by molecular signatures. For instance, 20% of a total of ~500 CRISPR loci identified in bacterial genomes were linked with assembled viral populations resulting in a highly interconnected network of viral-host historic integrations. Together these findings establish foundational empirical datasets for interpreting freshwater viral community dynamics and virus-host interactions and have implications for modeling natural microbial communities.

#30: Iterative subtractive binning of freshwater chronoseries metagenomes recovers nearly complete genomes from over four hundred novel species

Presenting Author: **Luis M. Rodriguez-R**
Email: lmrodriguezr@gmail.com

Complete Author List: Luis M. Rodriguez-R (1,2); Despina Tsementzi (1); Chengwei Luo (1); Konstantinos T. Konstantinidis (1,2)

(1) School of Civil and Environmental Engineering and
(2) Center for Bioinformatics & Computational Genomics, Georgia Tech, Altanta, GA.

Abstract:

The detailed study of natural microbial communities in most environments has been traditionally hindered by the limited number of genome sequences that can be recovered. The scarcity of reference material to identify microbial populations is therefore a major obstacle in current microbial ecology research across environmental and clinical settings, with a few exceptions such as communities with extremely low-diversity and cases where large collections of reference genomes exist (e.g., human gut microbiome). Recently, this issue has attracted special attention due to the availability of methods for the reconstruction of genomes from metagenomes, known as binning methods. A small number of large-scale have recently yielded collections representing novel deep-branching clades in the tree of life, mainly by applying single-sample binning. Here, we leverage a chronoseries consisting of 69 metagenomes from seven sites along the Chattahoochee River (Southeastern USA) including five lakes and two estuarine locations. We developed an iterative binning methodology gradually decreasing sample diversity in order to maximize the number and quality of recovered genomes. Our workflow consists of de novo sample clustering using hashed k-mer profiles, co-assembly and sub-assembly by sample group, binning, and genome quality evaluation. Once a group of high-quality genomes is identified, we filter out reads derived from that set through mapping, and iterate the procedure until no further significant gains in new genome sequences or phylogenetic diversity are observed. In the first iteration we recovered 199 high-quality genomes from 176 de-replicated groups (clades with Average Nucleotide Identity–ANI ≥ 95%). After eight iterations, we were able to increase this number to 1,126 genomes from 463 de-replicated groups, with a concomitant increase in explained community fraction from ~15% to 40-50%. Moreover, this collection represents a set of largely uncharacterized groups at high taxonomic ranks. Using the Microbial Genomes Atlas (MiGA), we were able to confidently classify only eight genomes (0.6%) at species or genus levels using the NCBI Genomes database as a reference. 216 genomes (19%) were classified at order level or below, and the majority of genomes in the set were classified only to class level (545 genomes or 48%). An additional 295 genomes (26%) potentially represent members of novel classes classified at phylum level, and 70 genomes (6.2%) potentially represent novel phyla. Finally, we were able to identify cases of seasonal rhythmicity and different levels of endemism in our collection, demonstrating quantitatively similar effects of seasonality and biogeography on freshwater microbial community assembly.

#31: Genome wide scan for signatures of social adaptation in Pseudomonas aeruginosa

Presenting Author: **Juan Castro**
Email: jcastro37@gatech.edu

Complete Author List: Juan C. Castro (1); Sam P. Brown (1)

(1) School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA.

Abstract:

The pathogen Pseudomonas aeruginosa exhibits complex social behaviors. Although the physiological basis of processes like quorum sensing its well understood, it is still unclear, how gene content differences, shape these processes across different environments and selective pressures. Machine learning algorithms have been used in an attempt to bridge genomic and experimental data. However different methods show different predictive power when assessing multiple traits, so it is evident that one approach may prove reliable in one case and flawed in another. We developed a stramlined architechture to integrate several methods in the evaluation of trait prediction, and tested data from 138 environmental strains of P. aeruginosa. Using several regression based learning algorithms, we were able to assign function to several hypotethical proteins in the Pseudomonas aeruginosa genome. And relate them to traits of social behavior such as biofilm formation and quorum sensing regulation (e.g. elastase production).

#32: Genome-wide mapping of ribonucleotide incorporation using ribose-seq and Ribose-Map

Presenting Author: **Alli Gombolay**
Email: agombolay3@gatech.edu

Complete Author List: Alli Gombolay (1); Fred Vannberg (1); Francesca Storici (1)

(1) School of Biological Sciences, Georgia Institute of Technology

Abstract:

Ribonucleoside monophosphates (rNMPs) are the most prevalent non-standard nucleotides found in genomic DNA. The main known mechanisms through which rNMPs become incorporated into DNA are via DNA polymerases during DNA replication or DNA repair synthesis. Once incorporated into DNA, the highly reactive 2'-hydroxyl group of rNMPs can attack the backbone of DNA leading to breakage of the DNA strand. When left unrepaired, rNMPs can wreak havoc on genome stability by causing DNA strand breaks, replication stress, and spontaneous mutagenesis. To understand the biological consequences of rNMPs and their role in the pathogenesis of human disease, we must profile the distribution of rNMPs in the genome. Determining where rNMPs frequently occur will allow us to identify how rNMPs cause genome instability. Recent advances in laboratory techniques and computational methods provide the unique opportunity to capture these non-standard nucleotides and map their locations in the genome. One of these techniques is ribose-seq (Koh et al. Nature Methods 2015). Ribose-seq harnesses the power of alkaline cleavage and tRNA ligase to capture rNMPs embedded in DNA. In contrast to other techniques, ribose-seq directly captures rNMPs embedded in DNA and may be applied to any cell type at any stage of the cell cycle. Of particular interest is the incorporation of rNMPs in diverse genomes, such as microbiomes and metagenomes. While much attention has focused on the incorporation of rNMPs in eukaryotic genomes, it would be helpful to compare the incorporation of rNMPs in different kingdoms of life. Achieving the full potential of ribose-seq is dependent upon computational methods tailored to analyzing this type of data. The Ribose-Map toolkit is a novel collection of user-friendly, open-source, and well-documented scripts developed to profile the incorporation of rNMPs captured via ribose-seq. Ribose-Map allows the user to determine the genomic coordinates of rNMPs, calculate nucleotide frequencies, locate rNMP genomic hotspots, and create publication-ready figures. While the presence of rNMPs in microbiomes and metagenomes has yet to be studied, the combination of ribose-seq and Ribose-Map would allow us to explore the effects of rNMPs on the stability of these diverse genomes.

#33: Illumina Sequencing of CCHF positive Serum Samples

Presenting Author: **Gvantsa Chanturia**
Email: gvantsa.chanturia@ncdc.ge

Complete Author List: Nato Kotaria (1); George Dzavashvili (1); Giorgi Babuadze (1); Babak Afrough (2); Roger Hewson (2); Gvantsa Chanturia (1)

(1) National Center for Disease Control and Public Health of Georgia, Lugar Center for Public Health Research
(2) National Infection Service, Public Health England Porton, Salisbury, SP4 0JG, United Kingdom

Abstract:

The first case of Crimean-Congo hemorrhagic fever (CCHF) in Georgia was registered in 2009. Since then, thirteen new cases were detected and confirmed in 2013 with the highest number of annual cases (30) so far registered in 2014. Sporadic cases of this dangerous and highly transmissible virus continue to emerge in Georgia and the National Center for Disease Control and Public Health (NCDC) conducts surveillance of this clinically important pathogen using serological and molecular approaches and increases knowledge of the disease among population living in the affected area. Next Generation Sequencing (NGS) capabilities are well established at the Molecular Epidemiology Laboratory of NCDC/Lugar Center for Public Health Research with the assistance of the genomics team at Los Alamos National Laboratory. We have developed two strategies (amplicon-based and metagenomic) for sequencing CCHF virus. In this work, we present the results of sequencing the CCHF virus circulating in Georgia using both amplicon and metagenomic approaches. The sequencing was performed on human patient serum samples. Following extraction of viral RNA and amplification, DNA and RNA library preparations were performed according to the NebNext protocols and sequenced on Illumina MiSeq. The data were analyzed using CLC-Bio software. After genome assembly, the complete S segment was mostly obtained using an amplicon sequencing approach. Near complete S and partial M and L segments were assembled from one metagenomic sample. This methodology was established at NCDC/Lugar Center and is now used for molecular and epidemiological surveillance of the virus in Georgia to improve our understanding of this pathogen in the country. This work enables molecular epidemiological monitoring of CCHF virus and will be useful to understand transmission chains leading to better disease control in the future.

#34: A virus or more in (nearly) every cell: ubiquitous virus-host interactions in extreme environments

Presenting Author: **Shengyun Peng**
Email: shengyun.peng@gatech.edu

Complete Author List: Jacob H. Munson-McGee (1*); Shengyun Peng (2*); Samantha Dewerff (3); Ramunas Stepanauskas (4); Rachel J. Whitaker (3); Joshua S. Weitz (2,5); Mark J. Young (1,6) *equal contribution

(1) Department of Microbiology and Immunology, Montana State University, Bozeman, Montana USA.
(2) School of Biological Sciences, Georgia Tech, Atlanta, Georgia USA.
(3) Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois USA.
(4) Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine, USA.
(5) School of Physics, Georgia Tech, Atlanta, Georgia USA.
(6) Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, Montana USA.

Abstract:

Virus activity influence the maintenance and structure of microbial communities in natural environments, including extreme environments such as hot springs. Previous studies have used metagenomics-based analyses to identify functional interactions between unculturable viruses and hosts. However, in large, diverse communities there is insufficient contextual information to identify the joint identity of virus and host pairs from short reads. Here, we report on an analysis of virus-host interactions derived from single-cells in hot springs within Yellowstone National Park. We do so by leveraging previous population mapping to identify viruses inside host cells, as well as the taxonomic identity of both viruses and hosts. Consistent with recent findings, we show evidence for virus presence in nearly every cell. Moreover, leveraging population mapping data, we are able to categorize the identified short viral sequences into a landscape of “viral partitions”. This contextual information allows us to identify a continuum of ‘promiscuity’ in the environment, i.e., the tendency for viruses to interact with more than one host cellular host type. Overall, combining single-cell and population level datasets provides new means to understand the ecology and evolution of viruses and their microbial hosts.

#35: K-shuff: a Novel Algorithm for Characterizing Structural and Compositional Diversity in Gene Libraries

Presenting Author: **William B. Whitman**
Email: whitman@uga.edu

Complete Author List: Kamlesh Jangid (1,2), Ming-Hung Kao (3), Aishwarya Lahamge (4), Mark A. Williams (5), Stephen L. Rathbun (6), William B. Whitman (1)

(1) Dept. of Microbiology, University of Georgia, Athens, GA, USA
(2) Microbial Culture Collection, National Centre for Cell Science, Savitribai Phule Pune University, Pune, Maharashtra, India
(3) School of Mathematical & Statistical Sciences, Arizona State University, Tempe, Arizona, USA
(4) Institute of Biotechnology and Bioinformatics, Savitribai Phule Pune University, Pune, Maharashtra, India
(5) College of Agriculture and Life Sciences, Virginia Polytechnic and State University, Blacksburg, Virginia, USA
(6) Dept. of Epidemiology & Biostatistics, University of Georgia, Athens, GA, USA

Abstract:

K-shuff is a new algorithm for comparing the similarity of gene sequence libraries, providing measures of the structural and compositional diversity as well as the significance of the differences between these measures. Inspired by Ripley's K-function for spatial point pattern analysis, the Intra K-function or IKF measures the structural diversity, including both the richness and overall similarity of the sequences, within a library. The Cross K-function or CKF measures the compositional diversity between gene libraries, reflecting both the number of OTUs shared as well as the overall similarity in OTUs. A Monte Carlo testing procedure then enables statistical evaluation of both the structural and compositional diversity between gene libraries. For 16S rRNA gene libraries from complex bacterial communities such as those found in seawater, salt marsh sediments, and soils, K-shuff yields reproducible estimates of structural and compositional diversity with libraries greater than 50 sequences. Similarly, for pyrosequencing libraries generated from a glacial retreat chronosequence and Illumina® libraries generated from US homes, K-shuff required at least 300 and 100 sequences per sample, respectively. Power analyses demonstrated that K-shuff is sensitive to small differences in the library composition but at the same time it maintained specificity in detecting only signifcant differences in both structure and composition of Illumina® libraries or Sanger libraries. The combination of this extra sensitivity with high specificity offered by K¬-shuff was highly useful in looking at differences at much deeper levels, such as within OTU members across different samples. This is especially useful when comparing communities that are compositionally very similar but functionally different. K-shuff will therefore prove beneficial for conventional microbiome analysis as well as specific hypothesis testing in such studies. For more details, see: PLOS ONE, 11(12), 22 pages. doi:10.1371/journal.pone.0167634.

#36: Genome ABundance Estimation software (GABE)

Presenting Author: **Jeong-Hyeon**
Email: jeochoi@gmail.com

Complete Author List: Jeong-Hyeon (1,2), Justin Choi (1,2), and Youngik Yang (1,2)

(1) National Marine Bio-Resources and Informatics Center
(2) National Marine Biodiversity Institute of Korea, 101-75, Jangsan-ro, Janghang-eup, Seochun-gun, Chungchungnam-do, Korea

Abstract:

Genome abundance estimation is a crucial task to understand a microbial community. It is often achieved by aligning NGS reads obtained from environmental samples to a set of known microbial reference sequences followed by calculating relative abundances of constituent microbes. Incorrect abundance estimation often occurs due to ambiguous read assignment. In case that a read has more than one best matches to multiple species (say, N hits), it is often practiced by counting N times or 1/N; however, neither approach is correct. This issue has been already addressed by Genome Relative Abundance using Mixture Models (GRAMMy). Though GRAMMy delivers a simple and elegant solution by solving the ambiguous read assignment using a mixture model, it is not directly applicable to widely accepted read mapping based approach and the implementation is rather difficult to modify. Therefore, we re-implemented and optimized the approach: GABE uses multi-threading for faster run-time, sparse matrix for less memory usage, and takes BWA read mapping as an input. In a benchmark set where short reads are randomly generated from 49 Staphylococcus aureus strains, the accumulate errors of both N counting and 1/N strategies reach to 50% while the accumulate error of our approach is negligible.

#37: Predicting the host-range of phage from abundance time-series

Presenting Author: **Ashley R. Coenen**
Email: ashley.coenen@gatech.edu

Complete Author List: Ashley R. Coenen and Joshua S. Weitz (1, 2)

(1) School of Physics, Georgia Institute of Technology
(2) School of Biological Sciences, Georgia Institute of Technology

Abstract:

Viruses of microbes, or phages, are an important part of microbial communities. In marine environments, phages are estimated to turn over 10 to 40 percent of microbes daily, contributing to microbial mortality and redirecting nutrient flow between trophic levels. Phages can also alter host metabolism and mobilize gene exchange.

With advances in high throughput sequencing and viral metagenomics, it is relatively easy to characterize which types of phages are present in a community.

However, it remains difficult to infer which microbes these phages can infect ("host range"). This difficulty remains, even as host-range identification is essential for understanding the ecological dynamics of microbial communities.

Methods for inferring host-range from metagenomic data fall broadly into two categories: sequence-based and dynamics-based. Sequence-based methods look for hallmarks such as sequence homology, sequence composition similarity, or CRISPR markers within viral and microbial genomes. Dynamics-based methods, on the other hand, look at changes in phage and microbial abundances over time to predict causal links.

Here, we benchmark two dynamics-based inference methods *in silico*. The first method uses correlations between abundance time-series to indicate phage-host interaction. Contrary to widespread use, our results suggest that correlation is a poor indicator of interaction when interactions are not already known *a priori* (Coenen and Weitz, bioRxiv, 2017). The second method extends recent work by using sparse linear regression on a discretized nonlinear mechanistic model of phage-host dynamics (Jover et al, Roy Soc, 2016). We find that, unlike the correlation-based method, model-based inference accurately recovers host-range and is robust to variation in network structure and life history traits.

Finally, we discuss the potential for enhancing dynamics-based inference with sequence-based information, for example by accounting for lysogenic phages.